Quality of service enhancement in VoIP network

Asosheh, Abbas

Quality of Service Enhancement in VoIP Network

By

Abbas Asosheh

Submitted at the King's College, University of London,
for the degree of ***Doctor of Philosophy***
Department of Electronic Engineering
August 2005

Abstract

Voice quality is one of the main problems that prevent the Internet telephony from competing with the traditional circuit-switched telephone. Though, the Internet telephony imposes certain new Quality of service (QoS) requirements which are commonly specified by bandwidth on demand, low end-to-end delay, low delay variation, acceptable error or loss rate without retransmission and codec (coder/decoder) quality.

Two novel application sender base error correction schemes to compensate for packet loss and mitigation the effect of delay jitter in the VoIP network will be introduced. The evaluation shows acceptable bandwidth efficiency and the loss tolerance improvement by 17% and 50%, respectively, in the voice codee base redundancy (VCBR) scheme and the second scheme, VCBR using backup channel (VCBRBC), can reduce the loss rate and average end-to-end delay compared to the single path VCBR scheme. The GLM is invoked to assess the improvement in the packet loss probability in this new method compared with and without redundancy.

A new loss model for analysis of real-time voice transmission over IP networks is proposed which can be used to analyze various techniques for enhancing the QoS in the IP telephony applications. The RED (Random Early Detection) technique and de-jitter buffer will be used for estimation of the queuing delay and tuning the packet loss in the receiver, respectively. The accuracy of the model is verified through simulation and analytical results for different traffic conditions, and it is shown that the model predicts

the overall voice packet loss rate (late and dropped) over the Internet on the BE condition with good precision.

Furthermore, we will focus our attention on providing adaptability to IP telephony applications through variable bit-rate coding and adaptive VCBRBC on the basis of estimation of the network condition using RED and de-jitter buffer length. It will be shown that over the Internet ὸn the BE condition, these new schemes can reduce the loss rate (late and dropped) and the network resource usage is improved compared to the usual scheme.

Table of Contents

Table of Figures

Glossary

ACELP	Algebraic Code Excited Linear Prediction
ADPCM	Adaptive Differential Pulse Code Modulation
ARPA's	Advanced Research Projects Agency's
avg	average queue size
VCBR	Voice Codec-Base Redundancy
AVoIP	Adaptive VoIP
BA	Behavior Aggregate
VCBRBC	VCBR using Backup Channel
BER	Bit-Error-Rate
CBS	Committed Burst Size
CELP	Code Excited Linear Prediction
CIR	Committed Information Rate
CODEC	CODer/dECoder
CS-ACELP	Conjugate Structure-Algebraic Code Excited Linear Prediction
D_{dec}	decoder Delay
delay-EDD	delay-Earliest-Due-Date
D_{enc}	encoder Delay
DiffServ	Differentiated Service
DMP	Data over Main Path
D_{pack}	packetization Delay

D_{pro}	propagation Delay
Dqave	mean queue waiting time
DRP	Data over Redundant Path
DS	Differentiated Service
DSCP	Differentiated Service Code Point
D_{ser}	serialization Delay
DSP	Digital Signal Processor
EBS	Excess Burst Size
EED	End to End Delay
EST	Expected Service Time
FCFS	First Come First Serve
FEC	Forward Error Correction
FECBC	FEC using Backup Channel
FIFO	First In First Out
FTP	File Transfer Protocol
GLM	Gilbert Loss Model
GSM	Global System for Mobile Communication
HTTP	Hyper Text Transport Protocol
i.i.d	independent and identically distributed
IETF	Internet Engineering Task Force
IntServ	Integrated Service
IP	Internet Protocol

ITU-T	International Telecommunication Union-Telecommunication
jitter-EDD	jitter-Earliest-Due-Date
LP	Linear Prediction
Mbone	Multicast backbone of the Internet
MF	Multi-Field
MIPs	Millions of Instructions Per second
MOS	Mean Opinion Scores
MPLS	Multi-Protocol Label Switching
MP-MLQ	Multi-Pulse Maximum Likelihood Quantization
NGI	Next Generation Internet
NS2	Network Simulator (version 2)
NVP	Network Voice Protocol
PBS	Peak Burst Size
PCM	Pulse Code Modulation
PDF	Probability Distribution Function
Pdf	Probability density function
P_{DL}	Probability of the Dropped Loss packets
P_{LL}	Late Loss Probability
PHBs	Per Hop Behaviors
PIR	Peak Information Rate
POD	Play Out Delay
PSNs	Packet Switched Networks

PSQM	Perceptual Speech Quality Measurement
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RED	Random Early Detection
RMD	Random Midpoint Displacement
RTCP	Real-time Control Transport Protocol
RTP	Real-time Transport Protocol
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SNR	Signal-to-Noise Ratio
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VBR	Variable Bit Rate
VC	Virtual Clock
VoIP	Voice over IP
WAN	Wide Area Network
WDM	Wavelength-Division Multiplexing
WFQ	Weighted Fair Queuing

Chapter 1- Introduction

1.1 Background

We live in a time of rapid technological change, especially in the field of telecommunications. The most common types of traffic handled are interactive data, generally transmitted in short bursts of a few characters; file transfer, involving the transmission of up to millions of characters (or bytes) between computers or mass storage systems; and increasingly, digital voice. Facsimiles, images and other types of traffic are being considered for transmission as well.

Voice transmission is still the most common mode of communication worldwide. It involves by far the largest investment in installed plant. The telephone networks developed to handle voice cover every part of the globe. All projections indicate that voice will continue to be the heaviest user of communication facilities worldwide. Once telephone networks become fully digitized, any kind of data, whether interactive data or due to computers communicating with each other, using digital voice, or images, could presumably traverse a network [TJ00].

One of the hottest topics in telecommunications today is the use of data networks and the Internet in particular, to transport voice and fax traffic that classically run at the circuit-switched which refers to switching a shared infrastructure from one dedicated use to another dedicated use. In these networks, which generally transmit voice or data, a private transmission path is established between any pair or group of users attempting to communicate and is held as long as transmission is required. Since many voice calls

consist of a "dead time" when neither party is speaking, there is a tremendous waste of resources in a circuit-switched network. On the other hand, since the circuit is entirely dedicated to that one call, there is no need for any of the techniques we will discuss later to provide what we call quality of service (QoS).

Some other networks use packet-switched technology, in which blocks of data called packets are transmitted from a source to a destination. Source and destination can be user terminals, computers, printers, or any other types of data communicating and/or data-handling devices. In this technology, packets from multiple users share the same distribution and transmission facilities.

Two modes of packet-switched data transmission are commonly distinguished. In one case, that of *virtual circuit* transmission, a path is first set up end to end through the network. User packets then traverse the network following the path chosen and arrive at the destination node in the sequence in which they were transmitted. They share link and switch facilities along the way, being stored at each intermediate node until ready to be read out or forwarded along the appropriate outgoing link. This method of transmission is also called *connection oriented* transmission. The second mode of transmission of packets is a *connectionless* one with individual *datagrams* moving between source and destination nodes. No initial connection is set up in this case. Datagrams are forwarded through the network on an individual basis. Routing at intermediate nodes is commonly based on the destination address of the datagram, which must be carried by each datagram. Datagrams are not necessarily guaranteed to arrive at the destination in the order of their transmission. In both modes of packet-switched transmission, virtual circuit (connection oriented) and datagram (connectionless), packets are queued at intermediate

points along the route between source and destination. This introduces a time delay during transmission that does not normally appear in circuit-switched transmission [Sch88].

In a packet-switched network, the network is shared between users and applications so a single user may have several applications running which use the local network but are switched on an individual packet-by-packet basis to their ultimate (and frequently different) destinations. Now voice, video and integrated data are coming together into a single network. It is referred as the architecture for voice, video and integrated data. The importance of digital and data network communications has greatly increased with the explosion of the Internet Protocol (IP) [XXL96]. The adoption of packet-switching and its merging with circuit-switching helps drive this communications migration. There are many reasons for this: pricing advantages due to improved resource utilization, seamless transitions between mono-media and multi-media communications, as well as between human-to-computer and interpersonal interactions.

To offer a credible alternative to traditional circuit-switched telephony, Voice over IP (VoIP) must offer the same reliability and voice quality. In other words, the current IP network is designed for traditional digital data transmission which mainly cares about the overall transmission throughput and reliability by employing the best-effort service model. However, the TCP/IP protocol and the best-effort service model are not suitable for real-time streams since they cannot provide any bandwidth or delay guarantees. Therefore internet telephony imposes certain new QoS requirements. It is stated that QoS should be guaranteed. Three different types of QoS guarantee can be distinguished: hard,

soft and best effort- where different levels of guarantee are used for different types of traffic.

However, several challenging problems need to be solved to achieve the same level of quality offered by the traditional PSTN in the context of a packet-switched data network. The first major issue is that of delay, which in a generic packet-switched network is neither bounded nor predictable; secondly, packets can be lost; and the last issue is that, transport of voice over data networks is generally made possible by speech compression techniques [SRM97]. Each speech coding algorithm implements a different trade-off between out put speech quality, algorithm delay, bit rate, computational complexity and robustness to background noise. The overall objective of voice transmission over IP networks is to find an array of technical solutions to these challenges that guarantee the desired level of perceptual quality under most network conditions.

1.2 Contributions of the Research

The main contributions of this research are:

1- We introduced a new error control algorithm combined with source coding to improve packet loss tolerance and bandwidth efficiency of VoIP network [AMS03].

2- A selective dropping mechanism is introduced for providing adaptive rate for IP telephony applications using variable bit-rate voice coding algorithms. The network conditions are measured in terms of packet delays and losses, using the average queue size in Random Early Detection (RED) as an effective mechanism to control the congestion in the IP network. The performance of the introduced

algorithm is verified by comparing the performance of simple drop-tail (FIFO) queuing and RED in a differentiated service enabled network [AS03].

3- A new voice codec-base redundancy (VCBR) scheme using a backup channel to send redundant information instead of piggy-backing, the main packet is proposed. The Gilbert loss model (GLM) is employed to verify the improvement of the packet loss probability [ASC04].

4- A new model is proposed for IP telephony applications. We introduce a complete loss model using system characteristics, such as queuing delay, delay jitter, and number of dropped packets. The accuracy of the model is verified through comparison of the analytical results with those obtained by numerical simulation.

5- An algorithm for adaptation of the transmitter and the receiver in the IP telephony applications is proposed. Play-out delay (de-jitter) time and number of independent transmission paths are used as two degrees of freedom in optimizing the QoS on the VoIP networks. The de-jitter buffer length is increased in response to the detection of early congestion through the RED method; whereas a backup channel for transmission of a redundant voice stream is employed to accommodate a reasonable de-jitter buffer length. We show that the overall average packet loss probability will be reduced by applying our proposed adaptive technique [AS04].

1.3 Organization of the Thesis

The thesis is organized as follows. Chapter 2 provides a literature survey and examines the specific characteristics associated with voice and data communications.

packet-switched networks, as well as connection-oriented and connectionless networking technologies. The main approaches to achieve QoS guarantees such as integrated service, differentiated service, different type of codecs and their effects on the level of QoS guarantee will be reviewed in the literature.

Chapter 3 provides an overview of traffic and channel modeling for telecommunications networks, such as self-similar traffic modeling and the Gilbert loss model; and queuing theory, which allows us to present the work in a probabilistic framework.

In Chapter 4, the VCBR and the Gilbert channel loss model are introduced. In addition we will show that the bandwidth efficiency increases at least by 13 percent and the tolerable packet loss can be increased to around 10 percent in this method. However, the total required capacity only increases by less than the codec bit rate. We propose a method that invokes the VCBR control algorithms, and a new error correction method which uses VCBR using the backup channel (VCBRBC) whilst employing a source codec to reduce the overall computational complexity. Finally, we evaluate the performance of the proposed error correction technique using the Network Simulator NS2 [FV05].

Chapter 5 presents a complete loss model for a VoIP network using the system characteristics of queuing delay, delay jitter, probability of dropped packets, and their effects on the network. We will give an estimate of the end-to-end delay and dropping probability voice packets over IP communication. The accuracy of the model is verified through simulation and analytical results for different traffic conditions, and it is shown

that the model predicts the overall voice packet loss rate (late and dropped) over the Internet on the BE condition with good precision.

In Chapter 6 we will focus our attention on adaptive IP telephony applications by introducing variable bit-rate coding algorithms, and use a selective dropping mechanism to make the deadline-misses evenly distributed during congestion. Moreover, an adaptive VCBRBC 'scheme is proposed on the basis of estimating the network conditions, measured in terms of the de-jitter buffer length. This is done by using the average queue size in in the context of RED to control the amount of redundancy or the need for making use of a backup channel, more efficiently. These algorithms aim to control the load of the network and use the network resources efficiently.

The conclusion and all my recommendations for future work will be outlined in Chapter 7.

Chapter 2- Literature Survey

2.1 Introduction

The growth of packet-based services is leading to the integration of voice, video and data over packet-switched networks. The importance of data communications has greatly increased with the explosive growth of the Internet. The adoption of packet-switching and its merging with circuit switching, helps drive this communications migration. Among the many reasons for this are the pricing advantages due to improved resource utilization; seamless transitions between mono-media and multi-media communications, and also between human-to-computer and interpersonal interactions [JCG04].

The slightly lower voice quality is one of the main problems that prevent Internet telephony from competing with the traditional circuit-switched telephone. The current Internet is designed for traditional data transmission, which mainly cares about the overall transmission throughput and reliability by employing the best-effort traffic model [Gil00]. However, the TCP/IP protocol and the best-effort service model are not suitable for real-time streams since they cannot provide any bandwidth or delay guarantees. Therefore internet telephony imposes certain new QoS requirements which are commonly specified by bandwidth on demand, low end-to-end delay, low delay variation, acceptable error or loss rate without retransmission, and codec (coder/decoder) quality [BCM01]. Thus the overall objective of voice transmission over IP network is to find an array of technical solutions to guarantee the desired level of perceptual quality under most network conditions [RAHB03].

The purpose of this chapter is to provide a general overview of networking and to examine the specific characteristics associated with voice and data communications, packet-switched networks, as well as connection-oriented and connectionless networking technologies. The main approaches to achieve QoS guarantees such as integrated service, differentiated service, different type of codec; and their effects on the level of QoS guarantee, will be seen in the literature. The rationale behind the overview is to give the reader some insight into the research problem domain.

2.2 Distributed Network Architecture

Internet telephony, defined as real-time voice or multimedia communications over packet-switched networks (PSNs), is far from a novelty. It dates back to the early days of the Internet. The Advanced Research Projects Agency's (ARPA's) Network Secure Communications project implemented an infrastructure for local and transmits real-time voice communications as early as December 1973. The key goal of Network Voice Protocol (NVP), was to demonstrate the feasibility of secure, high-quality, low-bandwidth and real-time two-party phone calls over PSNs.

Two sets of standards are emerging today for Internet telephony, the first from the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T), and the second from the Internet Engineering Task Force (IETF). Recommendation H.323 is the principal ITU-T standard for Internet telephony [Tho96]. It is an umbrella standard which refers to many other standards. It was first released in 1996 then subsequently in 1998 and 1999 [Tog99]. As it is customary in the industry, we use the term "H.323" to refer to the set of ITU-T standards for Internet telephony,

including recommendation H.323 itself. The Session Initiation Protocol (SIP) is the principal IETF standard for Internet telephony [Sch99, IETF01]. It allows the establishment, modification, and termination of multimedia calls. SIP relies on a host of internet protocols including the Real-time Transport Protocol (RTP), for data transport. The same applies to H.323, as shown by Figure 2-1 which depicts a simplified Internet telephony protocol stack. H.323 and SIP are often compared and contrasted with each other. The premise is that most existing packet telephony customers are currently running H.323.

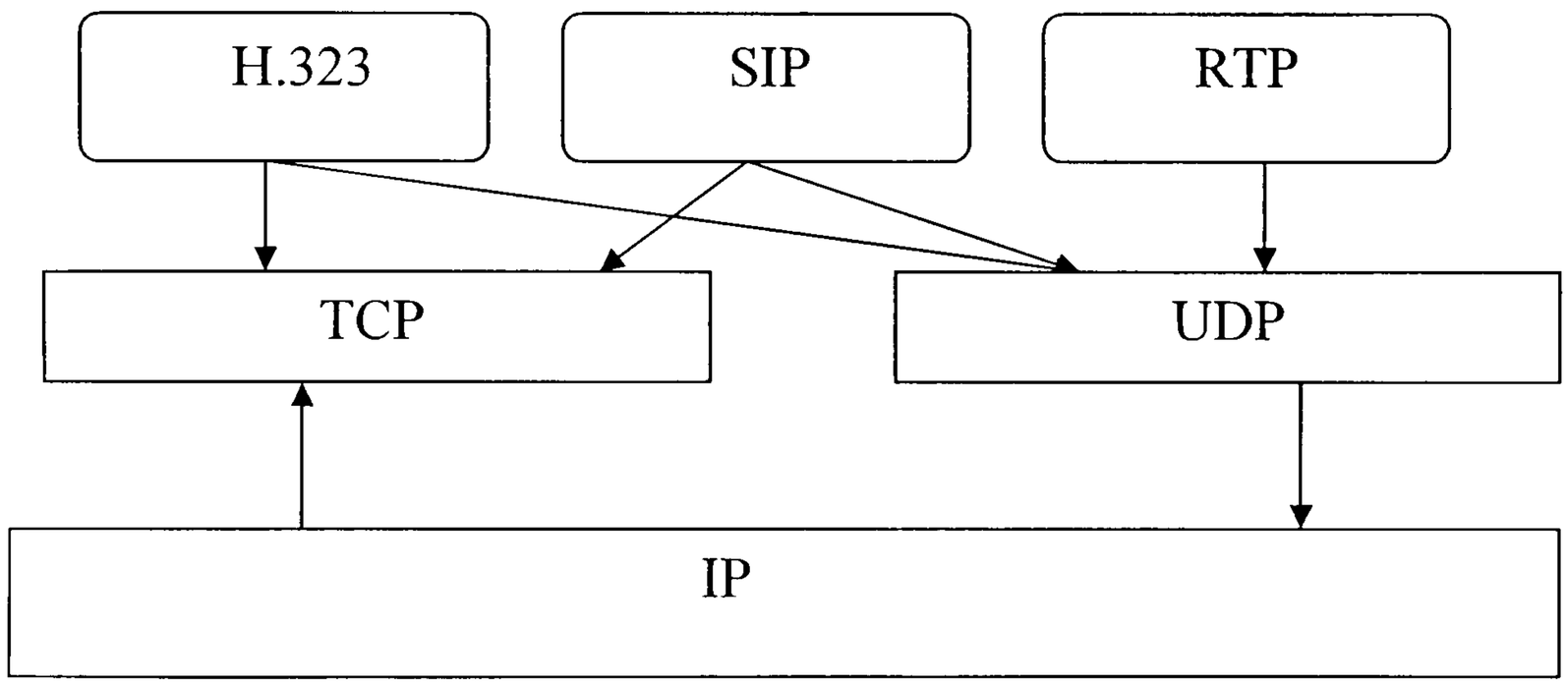


Figure 2-1: A Simplified Internet Telephony Protocol Stack. RTP: Real-time Transport Protocol, UDP: User Datagram Protocol, TCP: Transmission Control Protocol, IP: Internet Protocol.

Each protocol provides its own set of advantages and disadvantages within a packet voice network. It is possible to use both protocols within the same network, and it is definitely necessary to interconnect networks using one or the other. The H.323 protocol has been available for several years, and carriers have made a significant investment to build out many large H.323-based networks.

The Real-time Transport Protocol is a protocol developed by the IETF to allow transmission of continuous real-time information streams across IP-based networks [SFC96]. The Real-time Transport Protocol consists of two parts. Firstly, RTP defines the common RTP header format to be used with real-time data transmission; secondly, the Real-time Transport Control Protocol (RTCP) provides a mechanism for tracking and accounting information about the media stream itself and the quality of the underlying network, this is achieved by some low bandwidth information exchange in the background between sender(s) and receiver(s). Both protocols are carried in UDP datagrams [Bro00]. UDP packets are carried unreliably across an IP network: they may be lost, duplicated, and reordered. The transit delay of UDP packets is variable while capture and playback of real-time information streams typically is continuous. A sequence number and a timestamp in the RTP header allow receivers to determine the appropriate playback point for each information unit packet received, and thus preserve intra-stream timing. RTCP timestamps also allow correlation between different media streams to achieve inter-stream synchronization [SSR99].

2.3 Quality of Service (QoS)

One of the most outstanding examples of the integration support provided by IP is the Internet telephony or VoIP. Contrary to the switched telephone network, IP networks are not intended for transmitting voice. These new types of services include interactive multimedia communications involving digital audio and video. These services impose certain new QoS requirements on the Internet. QoS requirements are commonly specified by the following five parameters:

1. Bandwidth on demand;
2. Low end-to-end delay;
3. Low delay variation;
4. Acceptable error or loss rate without retransmission, as the delay would be unacceptable with retransmission.
5. codec (coder/decoder) quality— the audio quality produced by the encoding and decoding of analogue voice into digital code [Nol95].

The first four aspects/parameters are closely related. For example, when bandwidth on demand is truly met, end-to-end delay, delay jitter and error/loss rate will be low because packets don't have to wait excessively in queues and to be discarded. These aspects apply to all types of network traffic including data, voice and video. They are part of the design of any packet transport technology. But for voice, there is also a fifth element in the equation for providing service quality to the user; and this concern the quality of the audio codec [YSu01]. Although we generally state that QoS should be guaranteed, in practice, the user should be able to specify the degree (or level) of guarantees. In general, there are three levels of guarantees [VA99]:

- Hard or deterministic guarantee— user specified QoS should be met 100%. Previously, hard guarantees were achieved by reserving network resources based on the peak-bit rate of a stream.
- Soft or statistical guarantee— user-specified QoS should be met to a certain specified percentage. This is appropriate for continuous media because continuous media normally do not need 100% accuracy in playback. In addition, this type of guarantee uses system resources more efficiently.

- Best effort: no guarantee is provided and the application is executed with whatever resources are available. The traditional computer/network systems operate in this mode.

Different multimedia applications have different QoS requirements. We can classify multimedia applications into three categories, which we will all discuss, based on their general requirements. The first category is two-way conversational applications including telephone and videophone services. This category is characterized by its stringent requirement on end-to-end delay. An upper limit for one-way delay is 150 milliseconds according to the guidelines of recommendation G.114 of the ITU-T [ITU00]. This is end-to-end delay including total time taken to capture, digitize, encode/compress audio/video data; transport this from the source to the destination; and decode and display them to the user. The second category is broadcasting services where the source is live. The main difference from the conversational applications is that it is one-way communication and it can tolerate more delay. The third category is retrieval or on demand applications where the user requests some stored items and the server delivers them to the user. The main difference from the first two categories is that the media data has been stored and its characteristics can be studied in advance. In addition, this type of application can tolerate more delay than the conversational applications. The characteristics of these application types should be used in designing and implementing respective applications in order to provide required QoS while using network resources as efficiently as possible [WH01].

2.4 Main Approaches to QoS Guarantee

The above types of guarantee may all be required in a multimedia communication session. Different levels of guarantee are used for different types of traffic. It is up to the user to determine which type of guarantee to use. The charging policy is related to the level of guarantee. The hard guarantee is the most expensive and best effort is the cheapest. In some cases, one connection may use different levels of guarantee for different QoS parameters. For example, the user may request that the specified bit error rate should be met 100% (hard guarantee) but the specified delay jitter value should be met 90% (soft guarantee). The base Internet provides the BE service to applications and thus cannot meet the QoS requirements of multimedia communications [WS03]. However, many research and development efforts have been made towards providing QoS guarantee.

The simplified multimedia communication scenario is as follows. At the source, data is either captured live or retrieved from storage devices. This data is passed to the transport module where it is packetized and passed on to the Internet. The Internet is a connectionless packet network where each packet is routed independently from the source to the destination based on network addresses. At the destination, multimedia data are reassembled and passed to the application for playback of audio/video. We can summarize the main approaches to QoS guarantees as follows based on the above communication scenario [Fin02, RR02]:

- First—the speed of network links and routers is improved dramatically so that network congestion is very unlikely and QoS guarantees are provided automatically. A number of efforts are being made in this direction—this includes

all optical wavelength-division multiplexing (WDM) technologies, being investigated by the Next Generation Internet (NGI) initiative [THS03].

- Second—multimedia data should be coded in a way such that acceptable audio /video playback quality is still achieved in the event of packet loss or delay.
- Third—delay jitter must be removed at the destination, before data being played out. Delay jitter is caused by many factors, such as packet processing time differences, network access time differences, and queuing delay differences. This factor causing perceptual quality impairment on VoIP connection [HRR05]. Delay jitter can be removed with a first-in first-out (FIFO) buffer at the destination before playing. Arriving packets are placed in this buffer at a variable rate; the display device will remove samples in the buffer at a fixed interval determined by the nature of the medium. The principle of this buffering technique is to add a variable amount of delay to each packet so that overall delay for each packet from source to sink is the same. For this reason, it is often called delay-equalizing buffer or de-jittering.
- Fourth—communication architecture of the Internet is altered or improved to provide QoS guarantees. Three main architectures or models have been proposed:
    1. Integrated service (IntServ): IntServ is based on the idea of resource reservation. An appropriate amount of resources is reserved for each flow in order to meet its requirements [BRCDS94].
    2. Differentiated service (DiffServ): In DiffServ, all traffic is classified into certain number of classes, which are indicated by a special field. The

network then treats packets with different service classes differently to provide overall better performance.

3. Multi-protocol label switching (MPLS): MPLS can forward packets quickly and treat packets with different labels differently, once a label switched path is established. Thus it can be considered as a combination of IntServ and DiffServ.

- Fifth—it is a common requirement of multimedia communication to send data from one source to multiple destinations. Efficient multicasting protocols are needed to reduce bandwidth requirements [ROZY05].
- Finally, end systems (including servers and clients) must provide mechanisms to handle multimedia data efficiently and effectively in order to provide end-to-end QoS guarantees [HTW04].

In practice, all the above approaches should be implemented in order to achieve QoS guarantees effectively and efficiently. We now move onto describing the main principles and techniques of the second, third, fourth and final approaches.

2.5 Voice Codec

A codec (CODer / dECoder) is an algorithm, usually implemented in a DSP on a voice gateway or in software in a desktop computer, which converts between analogue voice and a digital representation of that voice. Coding is done at the entry point to a packet network and decoding is done upon egress from the packet network.

The real key to sending voice over any packet data network is digitization and compression, and VoIP is no exception. The three steps of digitization are[JC81] :

1. Sampling—samples are taken at a rate of 8000 times per second.
2. Quantization—each sample is quantified in comparison to a scale that has delineations grouped in segments.
3. Encoding—each quantified sample will produce an encoded 8-bit word that represents the sample's amplitude.

Compression offers several advantages, one of which is the reduction of raw bandwidth required to support the information transfer. There are two types of voice codec [WP01]:

- Waveform codecs (G.711, G.722 ,G.725, G.726, G.727)
- Vocoders (G.723x, G.729x)

Wave form compression is a subset of compression schemes, which include pulse code modulation (PCM) and its related derivations. This family of compression schemes is known as wave form coding because quantization and encoding (that is, sampling for quantity and assigning a binary value to the quantity) tracks and follows the actual analog wave form as it develops in real time. Coding techniques are standardized by the ITU, headquartered in Geneva, Switzerland. ITU-T standard G.711 codec is for digital voice delivery in the public switched telephone network (PSTN) and through PBXes. It is widely used in the telecommunications field because it improves the signal-to-noise ratio without increasing the amount of data. There are two subsets of the G.711 codec: Mu-Law and A-Law. Mu-Law is used in North American and Japanese phone networks, while A-Law is used in Europe and elsewhere around the world. Both Mu-Law and A-

Law use compressed speech carried in 8-bit samples. They use an 8-kHz sampling rate, and 64 Kbps storage.

The G.722 and G.725 use ADPCM (Adaptive Differential Pulse Code Modulation) with 8 bits per sample to improve speech quality beyond what is normally referred to as toll quality. They produce 64 Kbps, the same as G.711, but since the code is a more efficient representation of the speech, the quality is improved. They are useful in speaker phone or video conferencing situations in which improved audio quality is required. G.726 and G.727 use ADPCM to compress speech to 40, 32, 24, and 16 Kbps using respectively 5, 4, 3, or 2 bits per sample. In G.726, the sender and receiver must agree on which of these four coding methods to use; that method subsequently remains fixed for the duration of the call.

Unlike waveform coding, vocoding is based less on the analogue waveform than on the human vocal tract. Vocoders produce voice packets containing bits and bytes each of which represents some aspect of the voice code in a standard format [KM01]. This is in contrast to waveform codecs which produce a stream of bits/bytes without any packetization. It divides speech into voiced and unvoiced speech:

- Voiced speech resonates at a frequency that is characteristic of the human vocal tract, and is coded as the frequency plus the amplitude.
- Unvoiced speech is typically a consonant such as T or D and is represented as amplitude without any specific frequency.

The coder operates on speech frames of 10 ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For every 10 ms frame, the speech signal is analyzed to extract the parameters of the Code Excited Linear Prediction (CELP) model

(linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains). These parameters are encoded and transmitted. At the decoder, these parameters are used to retrieve the excitation and synthesis filter parameters. The speech is reconstructed by filtering this excitation through the short-term synthesis filter. The short-term synthesis filter is based on a 10th order Linear Prediction (LP) filter. The long-term, or pitch synthesis filter is implemented using the so-called adaptive-codebook approach. After computing the reconstructed speech, it is further enhanced by a post-filter.

Vocoding achieves a significantly lower bit rate than waveform coding and requires more processing power in the codec. G.723.1, G.728 and G.729 are vocoding standards from the ITU-T which achieve 5.3, 9.6, and 8 Kbps respectively, with a minimal degradation in quality compared to toll quality speech. The G.729x uses CS-ACELP (Conjugate Structure-Algebraic Code Excited Linear Prediction) and runs at 8 Kbps, a compression rate of 8:1. More recently, ITU-T has standardized two extensions of the 6.4 Kbit per second and 11.8 Kbit per second, indicated as G.729 annex D and E respectively [ITU99], [ITUG00]. There are, however, certain differences between them, i.e. the use of a reduced codebook and less fine quantization of some parameters such as pitch delay and gain. The basic scheme is the same as that of G.729 but there are some variations that do not only concern the quantization of the parameters.

Parameters	G.729	Annex D	Annex E forward	Annex E Backward
LPC	18	18	18	
Pitch period	13	12	13	13
Parity bit	1	0	1+1+1	1+1+1
Codebook	34	22	70	88
Pitch & code Book gain	14	12	14	14
TOTAL	80	64	118	118

Table 2-1: Bit allocation, every 10 ms, for G.729 and other operative modes.

The most important novelty is the introduction of backward linear prediction analysis, for better coding of music and speech gathered in the presence of stationary noise. The main body of G.729 and Annexes D and E provide a bit-exact, fixed-point specification of a CS-ACELP coder at 8 Kbit per second; and lower and higher bit-rate extension capability at 6.4 and 11.8 Kbit per second. These two extensions have been a significant reference point for the development of the hybrid multimode/multi-rate codec. The bit allocation for G.729 and its extension are summarized in Table 2-1 [EH99, RB01].

2.5.1 Voice Quality Measurement

Voice quality from a codec is measured using Mean Opinion Scores (MOS) that use a scale of 1-5 on which 'toll quality' is 4. The ITU-T recommends the measurement of voice quality using the five categories shown in Table 2-2 — a subjective methodology called the Mean Opinion Score (MOS). Test subjects are gathered into a lab environment and asked to rate voice quality through varying methods of compression. You must also

and asked to rate voice quality through varying methods of compression. You must also keep in mind that the ITU-T's recommendation for using MOS as a quality tool only specifies how to conduct the tests. The ITU-T itself does not publish individual MOS scores for individual codecs. Results will therefore vary from vendor to vendor and from test to test.

Although MOS is widely used, a newer method for measuring voice quality is being accepted by the industry. This method is called Perceptual Speech Quality Measurement (PSQM) and is assigned the ITU-T standard P.861 [ITU96]. PSQM was developed to measure voice quality in transmission systems originally developed for data, such as running voice over IP networks. The PSQM quality measurement can sometimes be a more precise measurement tool than MOS since it is not subjective and is sensitive to impairments seen on voice over data networks, such as delay and missing packets or frames. Some manufacturers of PSQM equipment include the capability to convert the PSQM result into MOS scores.

Quality Level	Description	Examples
Toll Quality (Grade 5)	Toll quality emulates and sounds like a copper wire. Exhibits similar quality to that of an analog end-to-end call with signal-to-noise ratios and harmonic distortions within acceptable limits.	Calls within a PBX from user to user or within a central office, such as calling a neighbor in the same geographic area.
Transparent Quality (Grade 4)	This is very similar to toll quality, with some tolerable distortions and almost imperceptible distractions. The distortions may be discernable to the most critical user, but are not annoying.	Calling long distance, from state to state, or between neighboring countries.
Conversational Quality (Grade 3)	Conversational quality has perceptible distortions and annoying distractions. In this category, the user begins to noticeably hear the degraded quality of the channel and may need to ask the speaker to repeat portions of the conversation.	Intercontinental calls or calls to third-world countries.
Synthetic Quality (Grade 2)	Synthetic quality is tolerable, but has very annoying distractions and poor reproduction of the speaker's voice fidelity. Because the reproduction is so poor, the listener hears what almost sounds like a machine-like re-creation.	Ship-to-shore telephony communications.
Unsatisfactory (Grade 1)	When the voice reproduction is this poor, the listener is sometimes forced to ask for a new channel. A simple exchange between the speaker and listener is strained to the point of the speaker repeating what's been said again.	Interference caused by malfunctioning equipment or induced from outside sources, such as radio interference.

Table 2-2: Five categories of the Mean Opinion Score (MOS).

Finally, table 2-3 shows the summary of the different important codecs and their essential parameters.

Codec	Bit Rate (Kbps)	Payload Size (Byte)	Frame Size (ms)	MIPs	Packetization Delay(ms)	MOS
G.711 PCM	64	160	20	.34	20	4.1
G.726 ADPCM	32	80	20	13	20	3.85
G.723.1 MP.ACELP	6.4	24	30	20	24	3.8
G.723.1 MP.MLQ	5.3	20	30	20	20	3.6
G.728 LD.CELP	16	5	2.5	33	5	3.61
G.729.A CS-CELP	8	10	10	10.5	20	3.92
G.729.E CS-ACELP	12	15	10	20	20	3.9

Table 2-3: Codecs and their important parameters. Millions of Instructions Per second (MIP), Mean Opinion Scores (MOS), Multi-Pulse Maximum Likelihood Quantization (MP-MLQ).

2.6 Accuracy

Methods of improving accuracy which introduce delay are unsuitable for real time interactive telephone conversations for which low latency is critical. However, they can be used in non-interactive streaming applications, such as voice messaging. Human interaction with a voice messaging system for sending or receiving voice mail does not need such a tight end-to-end delay as in the case of human-to-human interaction. As voice packets traverse a network, there are three situations that they can encounter which may cause information loss [PH98]:

- Buffer overflow—any packet-switched incorporates buffers that accumulate packets which are waiting to be switched or sent onto output ports. Due to the variability in the arrival rates of all the packets being handled by the switch, it

may be necessary to discard some packets. Although this is commonly referred to as 'buffer overflow', it does not necessarily mean that the buffer is physically full. Buffer management mechanisms can also discard low priority packets when a buffer is only partially full, in order to leave space in the buffer for subsequent packets that may arrive with higher priority. Voice packets are usually given high priority, but not necessarily the highest priority, which may be reserved for mission critical data or high resolution video.

- Laser malfunction—bit errors can be caused by temporary malfunction of the lasers and their control circuits that transmit light pulses along optical fibers. It is unusual for errors of this type to cause a single bit error. Instead a stream of consecutive bits is usually affected.
- Network failure—a network node such as a switch may fail due to electrical or optical component failure or due to power failure.

The accuracy can be improved by employing sender-based or receiver-based control mechanisms. A classification of sender-based mechanisms is shown in Figure 2-2 and receiver-based methods are shown in Figure 2-5 [Wri01].

2.6.1 Sender-Based Control Mechanisms

- **Retransmission** of error or missing packets is the most widely used sender-based recovery method. It is used in TCP for data traffic. A significant delay is introduced corresponding to the round-trip time of the request for retransmission to go from receiver to sender, and the retransmitted packet to travel back. Retransmissions are therefore suited to streaming as opposed to interactive voice applications. This

suited to non-interactive participants in a low error rate environment, and can be used for voice in applications such as training sessions and business presentations over packet networks— particularly IP, for which multicast is well developed and efficient.

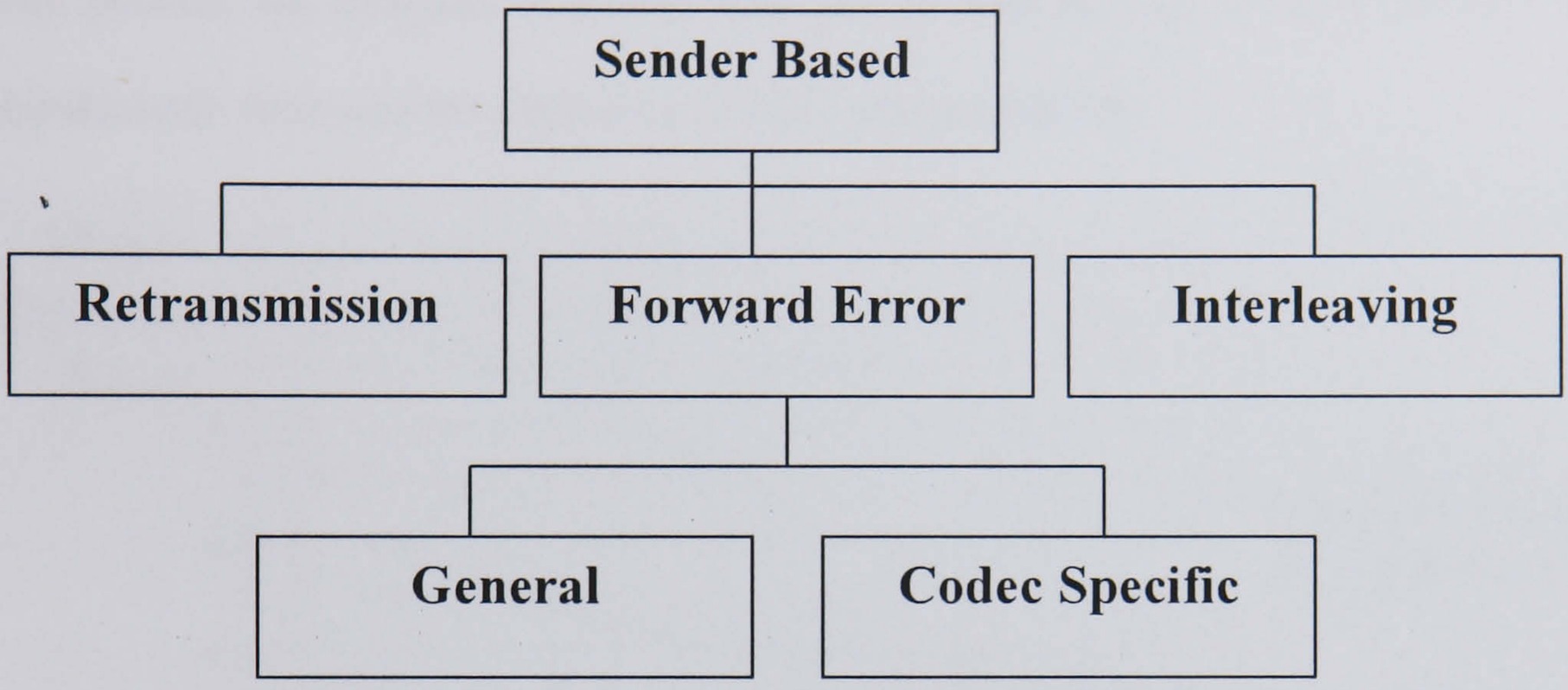


Figure 2-2: Sender based recovering lost packet methods.

- **Forward Error Correction (FEC)** can be used for voice applications in two ways [PHH98]. The first way is a general purpose technique that is applied after the voice codec has produced its voice packets, which in this method takes a block of voice packets and calculates a FEC code that is sent as a separate packet or packets. It introduces a delay dependent on the length of the block of voice packets. A missing or error packet anywhere within the block cannot be recovered until the FEC packet is received at the end of the block. This method is particularly suited to a high loss rate environment where retransmissions would cause a significant increase of traffic in the network. The second way is used as part of the voice codec itself. Codec-specific methods produce two versions of the voice code: the regular code and a more compressed code that can be used in case of errors in the regular code. In this method there is a slight degradation in speech quality at the receiver end, but not so significant a

degradation as if the packet were completely missing [AMS03]. Figure 2-3 shows the benefit of including redundant low bit rate voice code in foreign language tutorials over the Mbone. In many practical cases, consecutive packet losses are correlated, due to the way packets are dropped. A packet loss can be followed by a burst of loss, which significantly decreases the efficiency of FEC schemes[HS98] .

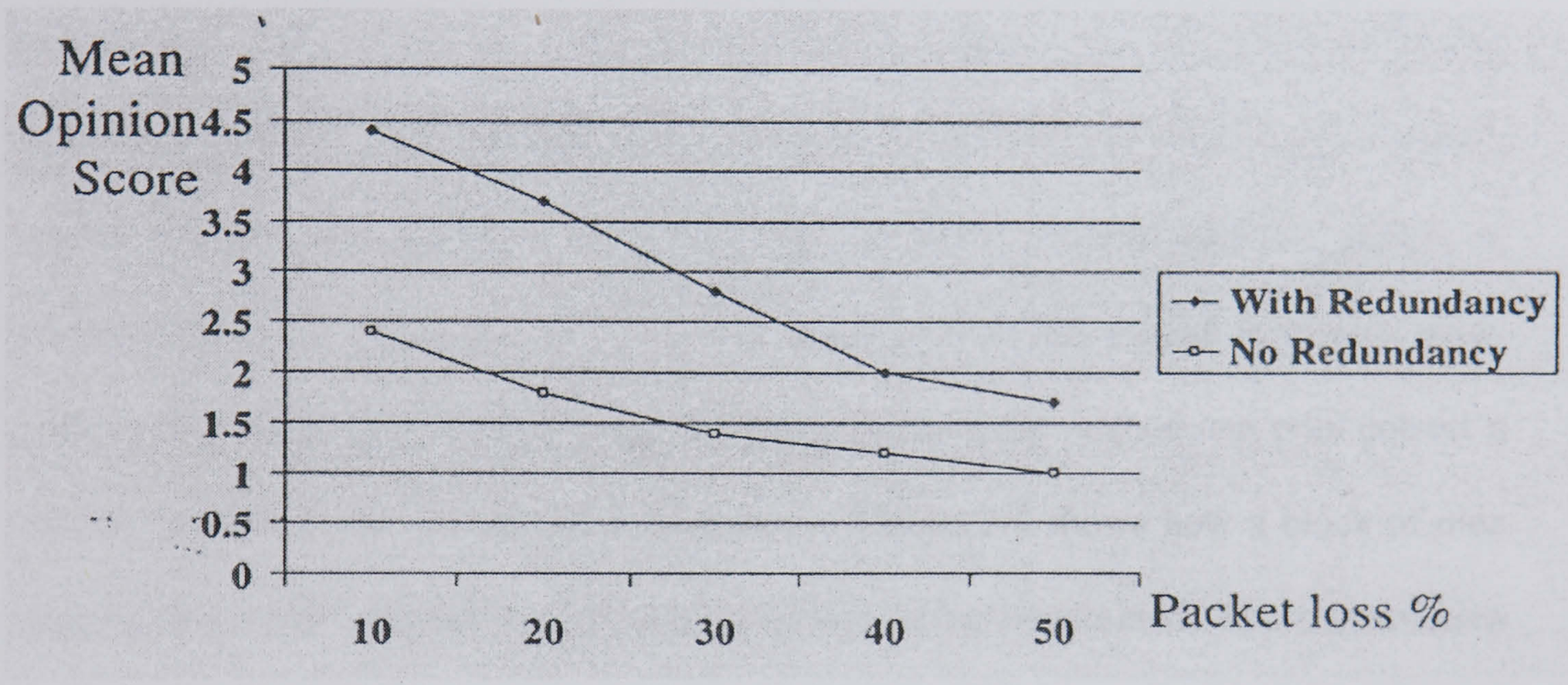


Figure 2-3: Benefit of data redundancy.

Recently, Path switching can potentially address the QoS guarantee that VoIP applications often require without requiring new network mechanisms, simply by leveraging the robustness to performance variations available from connectivity options such as multi-homing and overlays [TXEG05]. Overlay networks have emerged as a means to enhance end-to-end application performance and availability. Multi-path overlay transmission [RS04], attempt to leverage the inherent redundancy of the Internet's underlying routing infrastructure to detour packets along an alternate path when the given primary path becomes unavailable or suffers from congestion [HWJ05]. Backup channels as redundancy paths introduce the notion of availability to real-time

transmission at the cost of increasing the use of network resources. However, this over-provisioning of resources is potentially wasted, since fault rate is very low [GPM03, OC04].

- **Interleaving** is a method of spreading the effect of a group of consecutive errors over time. In situations of high loss rate, packet losses often occur in groups. This could be caused, for instance, by a congested buffer in a switch in the network. During congestion, many consecutive packets may be discarded; and when congestion abates, packet loss is rarer [LPD02]. FEC methods are often overwhelmed by groups of consecutive missing packets. The general purpose method has a limit as to how many packets in a block can be recovered, and the codec-specific method can only correct a missing packet if the next packet arrives intact; Figure 2-4 shows how a block of nine consecutive voice samples are interleaved before being transmitted. The transmission order is different from the original order of the voice samples, so that if a group of three consecutive samples is lost from the transmission sequence (2, 5 and 8), they are not consecutive after being un-interleaved at the destination. This makes it easier for one of the above methods, or for a receiver-based method, to be used to recover from the losses. Interleaving introduces delay dependent on the length of the block of voice samples or packets used, and is therefore suited to streaming voice applications.

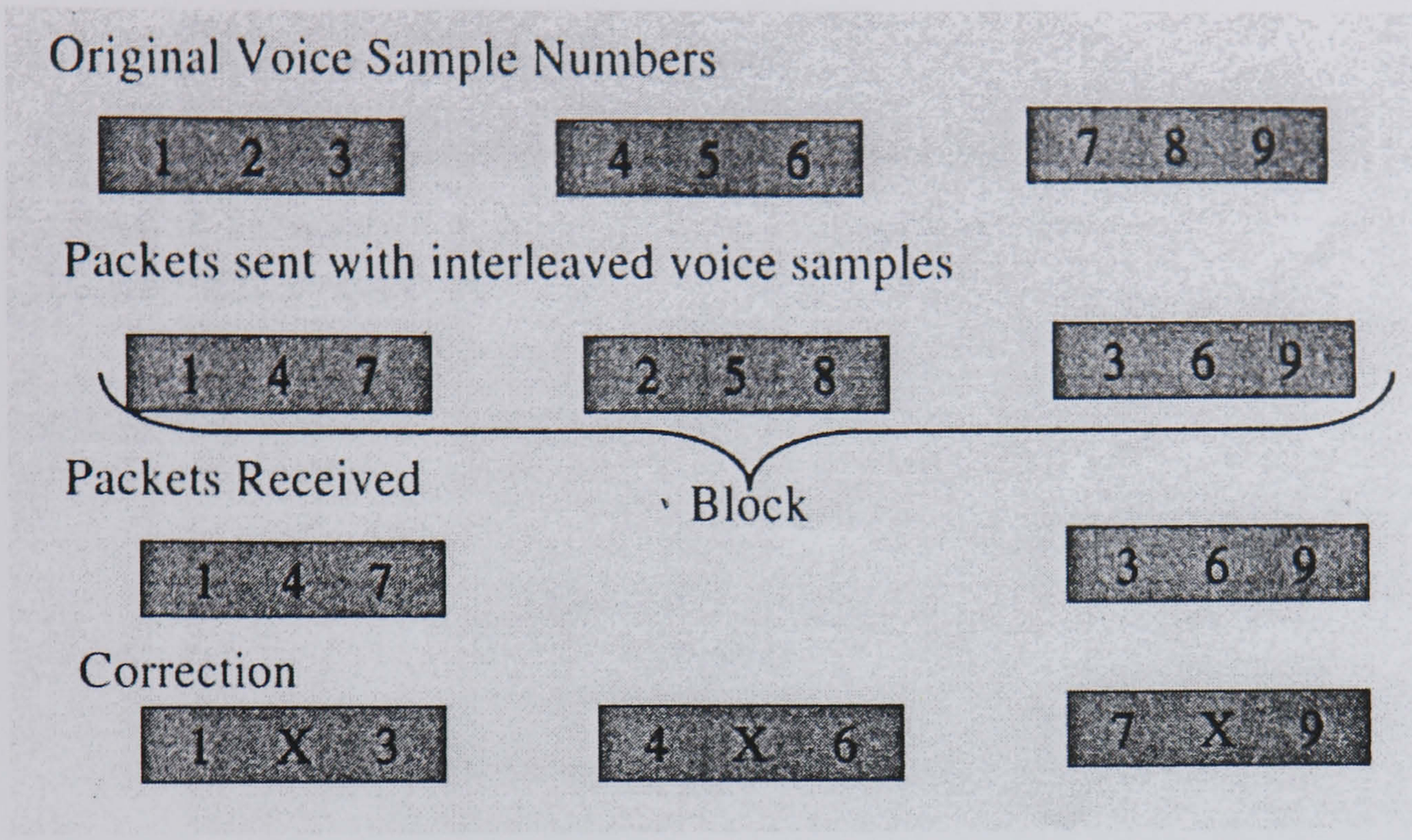


Figure 2-4: Interleaving method.

2.6.2 Receiver-Based Control Mechanisms

Receiver-based methods generally introduce less delay than sender-based methods and are therefore more suited to interactive human voice conversations. They do not recover lost packets as accurately as sender-based methods.

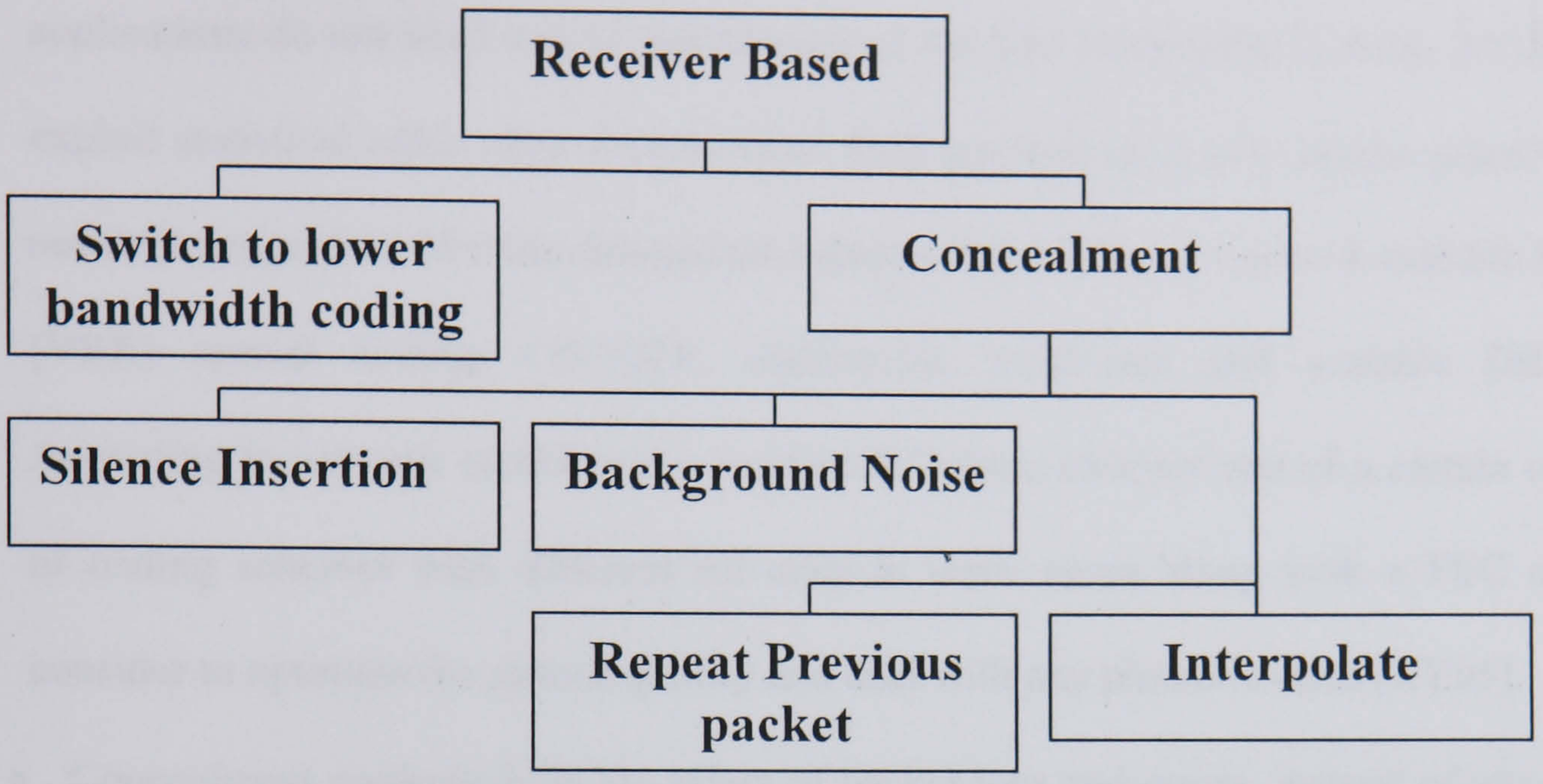


Figure 2-5: Receiver-based recovering lost packet methods.

They essentially trade accuracy for speed which is the right compromise for interactive speech. The alternative types of receiver-based methods are shown in Figure 2-5 and are now described.

- **Switching to Lower Bandwidth Encoding**: An IP multicast situation may be set up in which the receivers have a choice of different voice coding schemes [KB96]. A receiver may start by requesting a high quality coding scheme, which consumes correspondingly high bandwidth. Subsequently, the receiver may notice a high loss rate in arriving packets, which may be due to network congestion. Voice quality is dropping as a result of missing packets and may be worse than would be obtained from a lower bandwidth codec. It is therefore in the destination's interest to switch to a lower bandwidth codec. This switch also helps to alleviate network congestion, so that subsequently—after congestion has abated—the destination can switch back to the original higher quality codec. Adaptive VOIP (AVOIP) has several appealing features [BCDM01]. Firstly, an efficient use of network resources is granted since AVOIP

applications do not need a rigid partitioning of the link bandwidth; instead, AVOIP can exploit statistical rather than deterministic QoS guarantees; it also adapts gracefully to network congestion. We can distinguish between four different types of variable bit rate (VBR) speech coding: ON–OFF, multimode, multi-rate and scalable [BCR01]. According to network conditions, a 'multi-rate' codec chooses one of a certain number of coding schemes with different bit rates in some cases along with a FEC scheme consider to optimize the speech quality and deal with any phonetic class [KY05].

- Concealment methods hide the effect of packet loss and errors, instead of attempting to correct them [San98]. They are well suited to voice applications for which interactivity is more important than voice quality, and are incorporated in popular standard codecs such as G.723.l, G.728 and G.729. Concealment methods are based on the use of sequence numbers to detect whether a packet or voice sample is missing or delayed [PJAM04]. Figure 2-6 classifies five different concealment techniques according to how much processing is required to achieve a given improvement in voice quality.
    - The simplest method is to do nothing. i.e. to replace the missing packet by silence. This can be done with very short voice samples, up to a maximum of 5 ms, but thereafter it results in degradation in voice quality.
    - A significant improvement in voice quality at a minimal increase in processing can be achieved by inserting background noise instead of silence.

- A further improvement in voice quality can be achieved by replacing a missing packet by the previous packet instead of replacing it by background noise. Again only a minimal amount of processing is required.
- Global System for Mobile Communications (GSM) uses a method of repeating the previous packet for 20 ms and then fading the audio out over the next 320 milliseconds. This requires significantly more computation for a marginal improvement in voice quality and is suited to situations such as GSM where the packet length is long.
- Interpolation between the packets on either side of the missing one adds more to the computation requirement with only a very small improvement in voice quality and is not used very widely. Also it has the disadvantage of introducing a delay of one packet time.

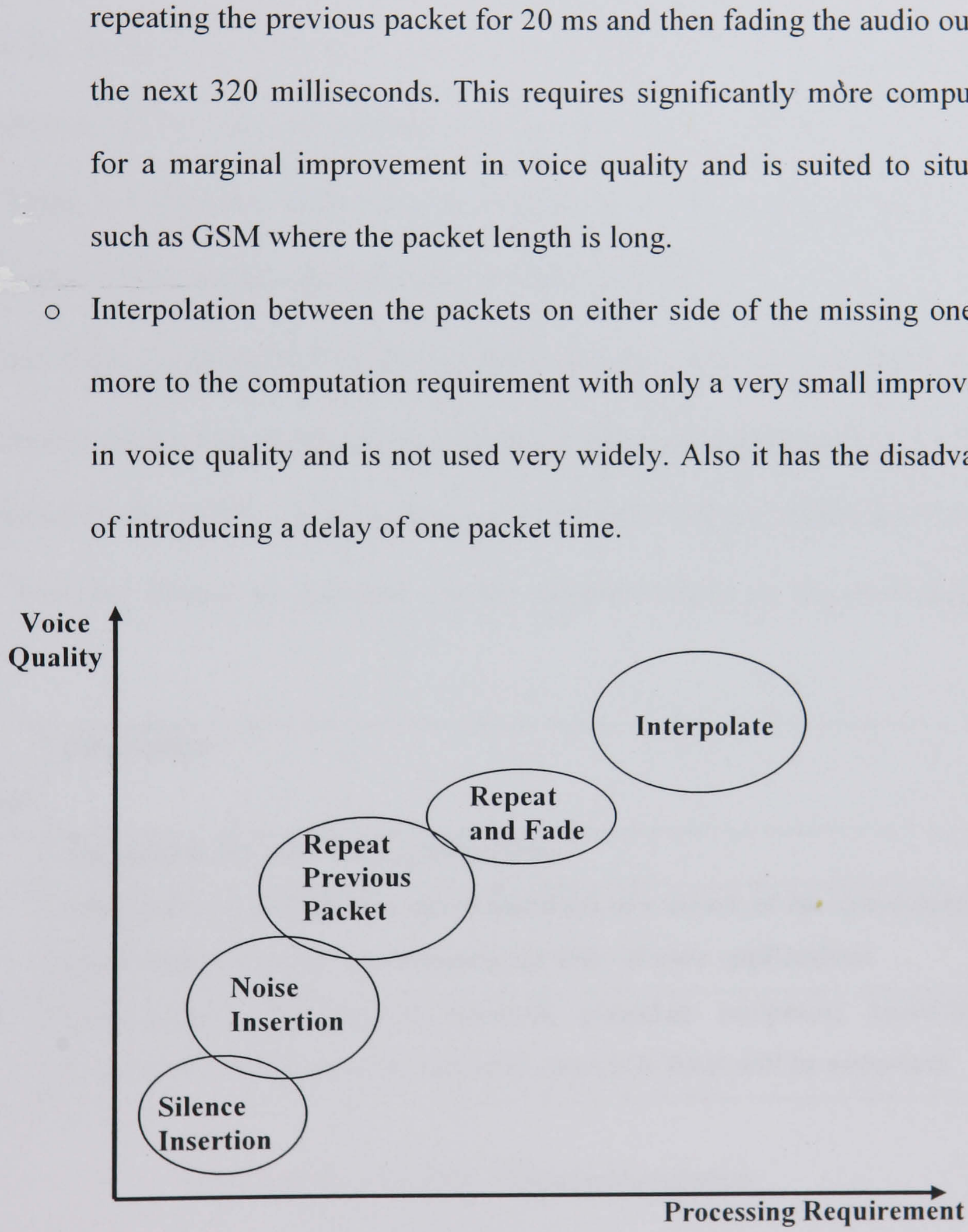


Figure 2-6: Concealment techniques according to processing requirement.

2.7 Latency

The majority of voice applications are interactive, for which latency is the primary QoS measure. Delay is measured end-to-end across the packet network from the point where the voice is coded at the source, across the packet network, to the point where it is decoded at the destination. The ITU-T considers network delay for voice applications in recommendation G.114. This recommendation defines three bands of one-way delay as shown in Table 2-4 [ITU00]. Delay can be caused in a voice communication by many different factors. There are two distinct types of delay [KT01]:

- Fixed delay components add directly to the overall delay on the connection.
- Variable delays arise from queuing delays in the egress trunk buffers on the serial port connected to the WAN. These buffers create variable delays, called jitter, across the network. Variable delays are handled via the de-jitter buffer at the receiving router/ gateway.

Range in Milliseconds	**Description**
0-150	**Acceptable for most user applications.**
150-400	**Acceptable provided that administrators are aware of the transmission time and its impact on the transmission quality of user applications.**
Above 400	**Unacceptable for general network planning purposes; however, it is recognized that in some exceptional cases this limit will be exceeded.**

Table 2-4: G.114, ITU-T Recommendation.

2.7.1 Voice processing delay

When Digital Signal Processor (DSP) chips are used to code voice, the processing time corresponds to the voice frame size. Coder delay is the time taken by the digital signal processor (DSP) to compress a block of PCM samples. Because different coders work in different ways, this delay varies with the voice coder used and processor speed. For example, algebraic code excited linear prediction (ACELP) algorithms work by analyzing a 10 ms block of PCM samples, and then compressing them. The compression time for a CS-ACELP process ranges from 2.5 ms to 10 ms depending on the loading of the DSP processor. If the DSP is fully loaded with four voice channels, the Coder delay will be 10ms. If the DSP is loaded with only one voice channel the Coder delay will be 2.5 ms. For design purposes we will use the worst case time of 10ms. Decompression time is roughly 10% of the compression time for each block [WP01].

However, because there may be multiple samples in each frame, the decompression time is proportional to the number of samples per frame. Consequently, the worst case decompression time for a frame with 3 samples is '3 x 1ms or 3ms'. Generally, two or three blocks of compressed G.729 output are put in one frame while one sample of compressed G.723.1 output is sent in a single frame. Best and worst case coder delays are shown in Table 2-5.

Coder	Rate (Kbps)	Required Sample Block (ms)	Best Case Coder Delay	Worst Case Coder Delay
ADPCM, G.726	32	10	10	10
CS-ACELP, G.729A	8.0	10	2.5	10
MP-MLQ, G.723.1	6.4	30	5	20
MP-ACELP, G.723.1	5.3	30	5	20

Table 2-5: Best and Worst Case Processing Delay.

2.7.2 Algorithmic Delay

The compression algorithm, which relies on known voice characteristics to correctly process sample block N, must have some knowledge of what's in block N+1 to accurately reproduce sample block N. This look ahead, which is really an additional delay, is called algorithmic delay and effectively increases the length of the compression block. Of course this happens repeatedly, such that block N+1 looks into block N+2, so forth and so on. The net effect is a 5 ms addition to the overall delay on the link. This means that the total time required to process a block of information is 10 ms with a 5 ms constant overhead factor.

- Algorithmic Delay for G.726 coders is 0 ms
- Algorithmic Delay for G.729 coders is 5 ms
- Algorithmic Delay for G.723.1 coders is 7.5 ms

Additionally, for simplicity, we will lump the coder delay, decompression delay, and algorithmic delay into one factor which we will call the coder delay. The equation used to generate the lumped Coder Delay Parameter is:

(Worst Case Compression Time Per Block)
+
(De-Compression Time Per Block)
X (Number of Blocks in Frame)
+
(Algorithmic Delay)

= "Lumped" Coder Delay Parameter

2.7.3 Packetization Delay

If a single voice packet is to be transported in a packet, there is minimal packetization delay. However, if $n > 1$ voice packets are to be packed into a single transport packet, there is a packetization delay of $(n-1)\ t$, where t is the packet time. At the source, the first of the n voice packets has to wait $(n-1)\ t$ while the other packets are being coded. At the destination the last of the n voice packets has to wait $(n-1)\ t$ while the earlier packets are being played out. Each byte is produced every 125 microsecond; therefore, the delay in putting n bytes into a packet is $(n-1)\ x125$ microseconds.

2.7.4 Serialization Delay

Serialization delay is the fixed delay required to clock a voice or data frame onto the network interface, and it is directly related to the clock rate on the trunk. Remember that at low clock speeds and small frame sizes the extra flag needed to separate frames is significant. Table 2-6 shows the serialization delay required for different frame sizes at different line speeds. This table uses total frame size, not payload size, for computation.

Frame Size (bytes)	Line Speed (Kbps)										
	19.2	**56**	**64**	**128**	**256**	**384**	**512**	**768**	**1024**	**1544**	**2048**
38	15.83	5.43	4.75	2.38	1.19	0.79	0.59	0.40	0.30	0.20	0.15
48	20.00	6.86	6.00	3.00	1.50	1.00	0.75	0.50	0.38	0.25	0.19
64	26.67	9.14	8.00	4.00	2.00	1.33	1.00	0.67	0.50	0.33	0.25
128	53.33	18.29	16.00	8.00	4.00	2.67	2.00	1.33	1.00	0.66	0.50
256	106.67	36.57	32.00	16.00	8.00	5.33	4.00	2.67	2.00	1.33	1.00
512	213.33	73.14	64.00	32.00	16.00	10.67	8.00	5.33	4.00	2.65	2.00
1024	426.67	146.29	128.00	64.00	32.00	21.33	16.00	10.67	8.00	5.31	4.00
1500	625.00	214.29	187.50	93.75	46.88	31.25	23.44	15.63	11.72	7.77	5.86
2048	853.33	292.57	256.00	128.00	64.00	42.67	32.00	21.33	16.00	10.61	8.00

Table 2-6: Serialization delay inm illiseconds for different frame sizes.

2.7.5 Queuing/Buffering Delay

After the compressed voice payload is built, a header is added and the frame is queued for transmission on the network connection. Because voice should have absolute priority in the router/gateway, a voice frame must only wait for either a data frame already playing out, or for other voice frames ahead of it. Essentially the voice frame is waiting for the serialization delay of any preceding frames in the output queue. Queuing delay is a variable delay and is dependent on the trunk speed and the state of the queue. Clearly there are random elements associated with the queuing delay.

2.7.6 Network delay

Although the network is often blamed for delay, particularly by novice users, it is only one factor in contributing to total delay. Network delay is highly variable. It includes propagation delay which depends on the end-to-end distance of the communication. For

instance the speed of light in fiber is about 200000 km per second, which introduces a delay of 15 ms on a 3000 km phone call. The other components in network delay are the switching delay, which is only a few microseconds; and the buffering delay in each switch, which depends on congestion. For instance, typical carrier delays for US frame relay connections are 40 ms fixed and 25 ms variable for a total worst case delay of 65 ms. For simplicity, we can include any low speed serialization delays in the 40 ms fixed delay. These are figures published by US frame relay carriers, to cover anywhere to anywhere coverage within the United States. It is to be expected that two locations which are geographically closer than the worst case will have better delay performance, but carriers normally document just the worst case.

2.7.7 De-jitter buffer delay

The buffering delay in network switches depends on other traffic volume and also on the priority of other traffic, and introduces considerable variability into total delay. The effect of the average of the jitter; the standard deviation of jitter on the IP network performances has been examined in [FNYF03]. In order to absorb the variability in the delay between one packet and another, a de-jitter buffer is implemented at the destination. When packets arrive, they are not played out immediately, but are kept in a buffer. When a sufficient supply of packets is available in the buffer, so that it is unlikely that it will run dry of packets due to variability in arrival times, we start to play out the voice to the destination user. We therefore absorb network jitter by introducing delay at the destination.

In other words the de-jitter buffer transforms the variable delay into a fixed delay by holding the first sample received for a period of time before playing it out. This holding

period is known as the initial play out delay. Proper handling of the de-jitter buffer is critical. If samples are held for too short a time, variations in delay may cause the buffer to under-run and cause gaps in the speech. If the sample is held for too long a time, the buffer can overrun, and the dropped packets again cause gaps in the speech. Lastly, if packets are held for too long a time, the overall delay on the connection may rise to unacceptable levels [JA05].

The initial play out delay is configurable, and the maximum depth of the buffer before it overflows is normally set to 1.5 or 2.0 times the total variable delay along the connection. The de-jitter buffer's actual contribution to delay is the initial play out delay of the de-jitter buffer plus the actual amount the first packet was buffered in the network. The worst case would be twice the de-jitter buffer initial delay (assuming the first packet through the network experienced only minimum buffering delay). In practice, over a number of network switch hops, it may not be necessary to assume the worst case.

2.8 Integrated service

A simplified Integrated Service (IntServ) model of a multimedia communications system is as follows [SP00, PH01]. An application specifies its QoS requirements, which are submitted to the system. The system determines whether it has sufficient resources to meet the requirements. If yes, it accepts the application and allocates the necessary resources to serve the application so that its requirements are satisfied. If it has insufficient resources to meet the application's requirement, it may either reject the application or suggest a lower QoS requirement that it can satisfy. In the latter case, if the application accepts the new set of QoS parameters, the application is accepted and

executed at the lower QoS. Failing this, the application is rejected, and it may try later in the hope that some resources may have been released by other applications [RLSG04]. Based on this simple operational model, the following elements are needed to provide QoS guarantees:

- A QoS specification mechanism for applications to specify their requirements.
- Admission control to determine whether the new application should be admitted without affecting the QoS of other ongoing applications.
- A QoS negotiation process so that as many applications as possible can be served.
- Resource allocation and scheduling to meet the QoS requirement of accepted applications [BF01].
- Traffic policing to make sure that applications generate the correct amount of data within the agreed specification.
- A QoS renegotiation mechanism is required so that applications can request changes in their initial QoS specifications.
- The actual QoS provided to the ongoing sessions should be monitored so that appropriate actions can be taken in case of any problems in providing specified QoS guarantees.
- Media scalability and graceful quality degradation techniques should be used.

2.8.1 QoS negotiation and renegotiation

When a connection with specified QoS is established, QoS parameters are translated and negotiated among all relevant subsystems. Only when all subsystems agree with and guarantee the specified QoS parameters can the end-to-end QoS requirements be met.

During the QoS negotiation process, a number of steps take place. First, QoS parameters are mapped or translated from one layer (or one subsystem) to another. Second, each layer or subsystem must determine whether it can support the required service; if so, certain resources are reserved for this session [NCP99]. Only when all subsystems accept the QoS parameters is the session established. Otherwise, the session is rejected. A sophisticated system may indicate to the user what level of QoS it can support. If the user is happy with the suggested quality level, the session is established.

Multimedia communications are normally not static. During an active communication session, changes in QoS may be necessary for various reasons. Therefore, it is necessary to provide QoS renegotiation mechanisms to meet the changing requirements of multimedia communications. It is sometimes not possible to meet the requirement to increase the QoS because the required resources may not be available [HMP00]. This issue is related to advance resource reservation. In advance resource reservation, users can reserve resources in advance by specifying the starting time and duration of the required session. This is very similar to conventional meeting room reservations. The acceptance of the reservation request is subject to the availability of requested resources at the required time. The important design issues of advance reservation schemes are how to divide the total amount of resources among sessions reserved in advance and sessions established on the spot (the conventional sessions); how to accurately specify reservation duration; and how to inform users who have successfully made reservations in the case of system failure.

There are two solutions to the first issue. First, the total resources (or capacity) are partitioned into two parts so that a portion is assigned to conventional sessions and the

other to sessions reserved in advance. The second is that all sessions share the total resources. The second issue is the most difficult to solve. When the duration is underestimated, the resources may not be available anymore before the session is successfully completed. But when the duration is overestimated, the resources may not be fully used.

2.8.2 Queue-scheduling disciplines

The most critical components inside a network that affect flow performances are switches/routers, where packets from different flows compete for the switching processing time and output link. A queue-scheduling discipline decides which packet is served next. Therefore queue-scheduling disciplines play an extremely important role in providing QoS guarantees to flows. Most queue scheduling disciplines rely on some sort of flow characterization. Admission control assumes that a certain type of queue-scheduling discipline is used when calculating resource requirements [YS01].

There are two types of queue disciplines: work conserving and non-work conserving. In work conserving scheduling schemes, the system may not be idle if there is data in the queue. An example of a work-conserving discipline is FIFO, in which a packet is always sent in each timeslot when the buffer is not empty. In a non-work conserving discipline each packet is assigned, either explicitly or implicitly, an eligibility time. The server idles (not transmitting any data) when no packets are eligible, although there are packets in the queue. Whether a service discipline is work conserving or non-work conserving affects buffer space requirements, delay, and delay jitter. Work-conserving disciplines need less buffer and cause shorter delays but cannot bound delay jitter tightly. The

opposite is true for non-work-conserving disciplines which need a larger buffer, cause longer delay, but can bound delay jitter tightly.

A number of queuing schemes have been proposed, including virtual clock (VC), weighted fair queuing (WFQ), delay-earliest-due-date (delay-EDD), jitter earliest-due-date (jitter-EDD), stop-and-go, hierarchical round robin, random early detection (RED) and FIFO [BGMT98].

2.8.3 Active queue management

Over the last decade, the flow and congestion control mechanisms of TCP have been used to adaptively control the rates of individual connections sharing IP network links. However, the performance of the TCP congestion control mechanism in networks that implement FIFO packet discard has some drawbacks, such as synchronization of flows, inequitable distribution of packet loss among flows, and low utilization of network resources. Therefore, even with end systems equipped with important algorithms such as the TCP congestion avoidance, slow start, fast retransmit, and fast recovery mechanisms, the performance of the TCP congestion control algorithm over current drop networks can still be unsatisfactory.

Active queue management has been recommended by the Internet Engineering Task Force (IEFF) as a way of mitigating the above stated performance limitations of TCP over drop-tail networks [FJ93]. Random Early Detection (RED) is the first active queue management algorithm proposed for deployment in TCP/IP networks. The basic idea behind an active queue management algorithm is to convey congestion notification early to the TCP end-point so that they can reduce their transmission rates before queue

overflow and sustained packet loss occur. It is now widely accepted that a RED controlled queue performs better than a FIFO. However, RED has some parameter tuning issues that need to be carefully addressed for it to give good performance under different network scenarios. The RED algorithm detects congestion and measures the traffic load level in the queue using the average queue size [ZY02]. This is calculated using an exponentially weighted moving average filter and can be expressed as

$$avg_n = (1-w_q)\, avg_{n-1} + w_q . q$$

where q is the actual queue size, w_q is the filter weight which determines the time constant of the low-pass filter, and avg_n denotes the estimate of the average queue size at step n. When the average queue size is smaller than a minimum threshold (min_{th}), no packets are dropped. When the average queue size exceeds the minimum threshold, the router randomly drops arriving packets with a given drop probability. The probability that a packet arriving at the queue is dropped depends on the average queue length, the time elapsed since the last packet was dropped, and the maximum drop probability parameter (max_p). If the average queue size is larger than a maximum threshold (max_{th}) all arriving packets are dropped. As ***avg*** varies from min_{th} to max_{th}, the packet-marking probability P_b varies linearly from 0 to max_p.

$$P_b = max_P . (avg - min_{th})/(max_{th} - min_{th}).$$

One of RED's main goals is to use this combination of queue length averaging (which accommodates bursty traffic) and early congestion notification (which reduces the average queue length) to simultaneously achieve low average queuing delay and high throughput.

2.8.4 Strengths and weaknesses of the IntServ

The main advantages of the IntServ model is that QoS requirements of a session are guaranteed once resources are reserved successfully for the session. The challenge is how to provide QoS guarantee without reserving resources based on peak-bit rate; that is QoS is guaranteed while resources are shared/used efficiently. There are two approaches to achieving this. The first is to make use of statistical multiplexing to achieve soft QoS guarantees. The second approach is to properly characterize traffic so that hard QoS guarantees are achieved.

The main weakness of the IntServ model is that it [HTW04]requires a substantial amount of state information to maintain the reservation in each router along the path. This information is required to identify the flow, track the flow's resource consumption, police traffic and schedule the traffic based on the reservation commitment. The amount of state information may be acceptable at the edge of the Internet where the number of connections is small. But it will not be acceptable at the core of the Internet where millions of connections pass through. For this reason, the IntServ model is not scalable. Another associated problem is that the IntServ cannot be deployed progressively, as all routers have to reserve appropriate resources in order to maintain end-to-end QoS.

2.9 Differentiated services

The main problem of the currently implemented best-effort model is that all packets are treated the same although different types of service can be specified in the IPv4 header; while the main problem of the IntServ is that there are potentially infinite number of different types of traffic so each router has to store necessary information in order to

provide QoS guarantees to each type of traffic [DSR02]. There is a clear need for relatively simple and coarse methods of providing differentiated classes of service for Internet traffic, to support various types of applications, and specific business requirements. The Differentiated Service (DS) approach to providing quality of service in networks employs a small, well-defined set of building blocks from which a variety of aggregate behaviors may be built. DSs take a middle ground between the best-effort service and IntServ [DS99, Hao02]. It defines a fixed number of packet classes. All traffic types/packets are aggregated into these classes and the network/routers provide different services to different packet classes [WCHLT04].

2.9.1 Service classification

IPv4 has an under-used type-of-service byte in its header. The newer IPv6 has a header byte called traffic class as shown in Figure 2-7. In DS, the type-of-service (traffic class) byte is redefined as a differentiated service field [THS03]. The first six bits of that field is called DS Code Point (DSCP) which indicates the behavior each router is required to apply to the individual packet [RP00]. Packets with the DS code point set to zero receive the same service as they get today (the best effort service). Values between one and seven are defined to be backward compatible with the original IP precedence mechanism, to ensure that DiffServ technology can be deployed in the operational Internet progressively.

The DS filed can be assigned by the customer (the transmitter process) to indicate the desired service. Alternatively, the ingress router marks the DS field based on multi-field (MF) classification. MF classification classifies packets based on the contents of multiple

fields such as source address, destination address, type-of-service byte, protocols ID, source port number and destination port number. When a packet moves from one Internet Service Provider (domain) to another, it may be reclassified. Many service classes can be defined. The IETF working group has defined the following two new services (in addition to the best-effort service).

- **Expedited** or premium service: to provide virtual leased line service to applications requiring low delay and low jitter [JNP99].
- **Assured** service: to provide better than best-effort services to applications [Rez99,BWW99].

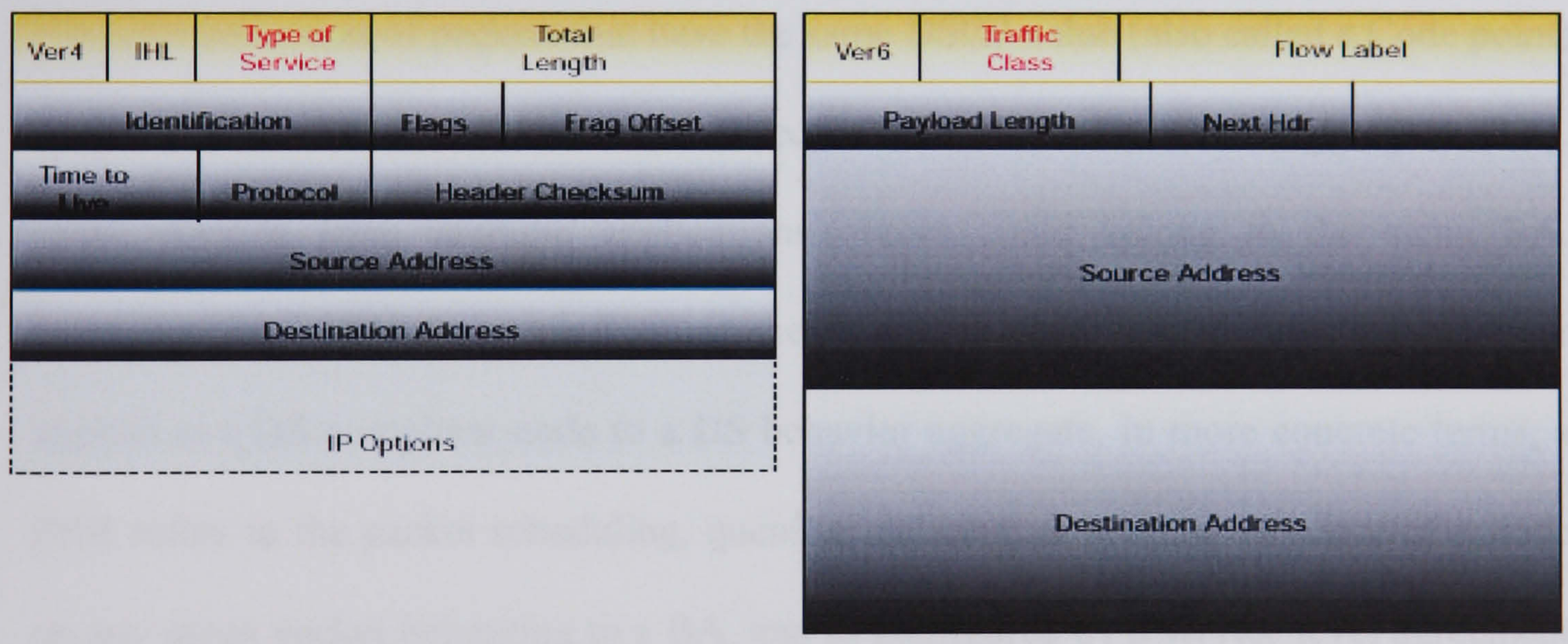


Figure 2-7: IPv4 and IPv6 Headers.

DiffServ is a priority scheme. Packets with different DS fields (class types) are treated with different priorities by routers. In order to deliver end-to-end QoS, this architecture [Gro02] has two major components:

- **Packet Marking**—unlike the IP-Precedence solution, the ToS byte is completely redefined (Figure 2-8). Six bits are now used to classify packets. The field is now called the DS Field, with two of the bits unused (RFC-2474). The 6 bits replace the three IP-

Precedence bits, and is called the DSCP. With DSCP, in any given node, up to 64 different aggregates/classes can be supported. All classification and QoS revolves around the DSCP in the DiffServ model.

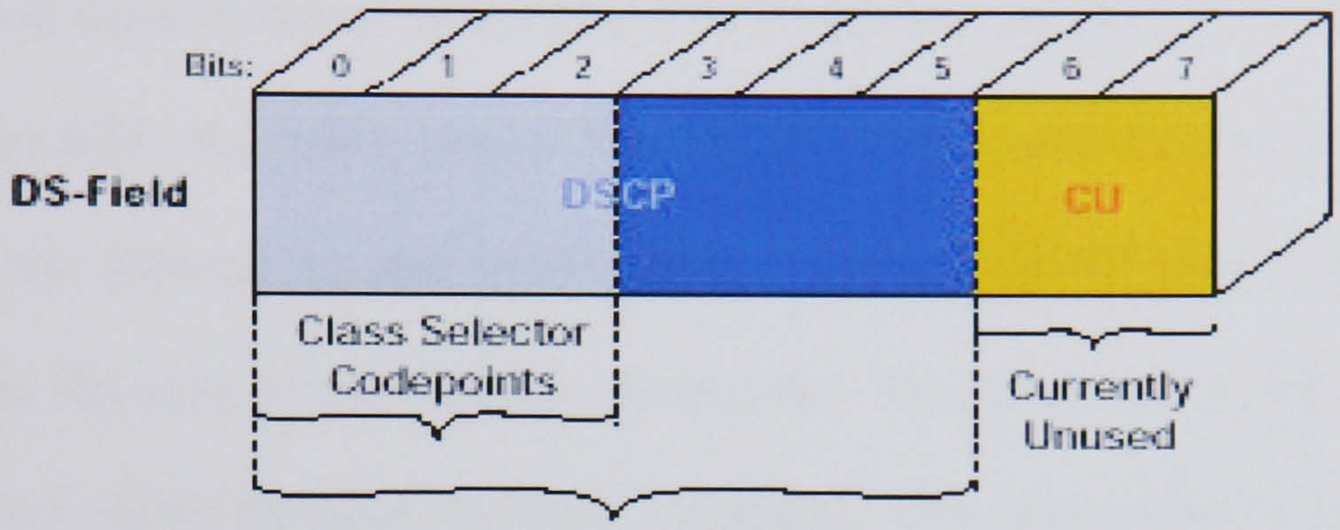


Figure 2-8: DiffServ Code point Field.

- **Per Hop Behaviors (PHBs)**—now that packets can be marked using the DSCP, how do we provide meaningful classification on flows, and provide the QoS that is needed? First, the collection of packets that have the same DSCP value (also called a Code point) in them, and crossing in a particular direction is called a Behavior Aggregate (BA). Thus, packets from multiple applications/sources could belong to the same BA. Formally, RFC-2475 defines a PHB as the externally observable forwarding behavior applied at a DS-compliant node to a DS behavior aggregate. In more concrete terms, a PHB refers to the packet scheduling, queuing, policing, or shaping behavior of a node on any given packet belonging to a BA, and as configured by a service level agreement (SLA) or policy.

2.9.2 Selective dropping mechanism and Adaptive Voice over IP

In AVoIP systems the rate of the single speech sources is dynamically adapted to the workload conditions. A variable bit rate (VBR) speech coder chooses the most appropriate bit rate from a predefined set of operating modes: source or network-driven

[GB97]. To guarantee a certain QoS even in critical conditions featuring great delays and background noise, it is necessary to control the peak rate, and therefore use a multi-rate codec, as mentioned in Section 2-3. The rate control algorithm which is described here, starting from loss measurements relayed by RTCP reports, and then by using the average queue size of the RED algorithm, tries to regulate the output rate of several voice sources. Rather, it uses the information that can be carried by cyclic RTCP receiver reports to let the source know the state of the ongoing connection. The adaptive algorithm follows the 'additive increase, multiplicative decrease paradigm.' The basic idea is that the source coder should reduce its rate when packet delays have been observed to have increased considerably above a 'high-mark threshold.' Also, it should drastically decrease the rate to, say, the minimum of its value, when severe congestion is detected (i.e. the packets are lost). Besides, it should switch to higher rates if no packets are lost and if the delay decreases below a 'low mark threshold.' This solution can also be used together with other QoS providing mechanisms, such as Diffserv or priority queuing, in which the network does not micro-manage its resources [AS03].

In the DiffServ model, packets are marked differently to create several packet classes. The core routers only classify packets based on the packet class instead of the individual micro-flow. Core routers do not need to process per-flow signaling or resource reservation. In DS, all the complexities are pushed out to the edge routers and the core routers are maintained as simple as possible. The differentiated services architecture is based on a simple model where the traffic entering a network is classified and possibly conditioned at the boundaries of the network, and then assigned to different behavior aggregates. In the approach taken by DS, individual micro-flows are classified at the edge

routers in the network into one of the many classes [SKS01]. It then applies a per class service in the core of the network. The core routers that forward the packet examine this marking and use it to decide how the packet should be treated. Most of the work in this scheme is done at the edge routers [Mup00]. These routers are responsible for classifying, using a multi-field classifier and traffic meter, and decide the next action to be taken on the packet. So it is relatively easier to be implemented in the Internet and has better scalability.

In the DS model, predefined policies are used to allocate resources to a particular client. These policies are based on the clients' peak traffic rate, the time for which the service is required, and the acceptable delay and jitter. A policy is established between a source and destination node. All flows matching that source-destination pair are treated as a single traffic aggregate. Policy for different traffic aggregate has an associated policier type, meter type, and initial code point. When a packet arrives at an edge router, it is examined to determine which aggregate it belongs to. The meter specified by the corresponding policy is invoked to update all state variables. A policy-enabled DiffServ architecture to provide predictable and measurable quality of service (QoS) for VoIP studied in [BB04]. The policier is invoked to determine how to mark the packet depending on the aggregate's state variables: the specified initial code point or a downgraded code point. Then the packet is enqueued accordingly. There are several different policy models, some of which are defined here [FV05]:

1. Time Sliding Window with 2 Color Marking (TSW2CMPolicer)—uses committed information rate (CIR) and two drop precedence. The lower precedence is used probabilistically when the CIR is exceeded.

2. Time Sliding Window with 3 Color Marking (TSW3CMPolicer)—uses a CIR, peak information rate (PIR), and three drop precedence. The medium drop precedence is used probabilistically when the CIR is exceeded and the lowest drop precedence is used probabilistically when the PIR is exceeded.
3. Token Bucket—uses a CIR and committed burst size (CBS) and two drop precedence. An arriving packet is marked with the lower precedence if and only if it is larger than the token bucket.
4. Single Rate Three Color Marker (srTCMPolicer)—uses a CIR, CBS, and an excess burst size (EBS) to choose from three drop precedence.
5. Two Rate Three Color Marker (trTCMPolicer)—uses a CIR, CBS, PIR, and a peak burst size (PBS) to choose from three drop precedence.

Chapter 3- Network Model

3.1 Introduction

In telecommunications, the term traffic or teletraffic implies the flow of information, (or data), in telecommunications networks of all kinds. From the first step as an analog signal carrying encoded voice over a dedicated wire or 'circuit' traffic now covers information of all kinds, including voice, video, text, telemetry, and real-time versions of each; including distributed gaming [Sch88]. Instead of the dedicated circuits of traditional telephone networks, packet switching technology is now used to carry traffic of all types in a uniform format (to a first approximation) as a stream of packets; each containing a header with networking information and a payload of bytes of 'data'.

A key concept in networking is the existence of network protocols, and their encapsulation. For instance, the IP is used to allow the transport of packets over heterogeneous networks. The protocol understands and knows how to process information, such as addressing details contained in the header of IP packets. However, by itself IP is only a forwarding mechanism without any guarantee of successful delivery. At the next higher level, the TCP provides such a guarantee by establishing a virtual connection between two end points and monitoring the safe arrival of IP packets; and managing the retransmission of any lost packets. On a still higher level, web-page transfers occur via the Hyper Text Transport Protocol (HTTP), which uses TCP for reliable transfer.

The resulting encapsulation 'HTTP over TCP over IP' therefore means that HTTP oversees the transfer of text and images, while the actual data files are handed over to TCP for reliable transfer. TCP chops the data into datagrams (packets) which are handed to IP for proper routing through the network. This organization offers hierarchal structuring of network functionality and traffic but also adds complexity: each level has its own dynamics and mechanisms, as well as time scales.

Over this landscape flows the teletraffic, which has even more levels of complexity than the underlying network. Three general categories can be distinguished [ABFRV02].

- Geographic complexity plays a major role.
- Offered traffic complexity relates to the multilayered nature of traffic demands.
- Temporal complexity is omnipresent. All of the above aspects of traffic are time varying and take place over a very wide range of time-scales; from microseconds for protocols acting on packets at the local area network level; through daily and weekly cycles; up to the evolution of the phenomena themselves over months and years.

As these new communications services evolve and the needs of users change, the enterprise must respond by modifying existing communications systems or by implementing entirely new ones. To this end, telecommunications professionals are being called upon to design and manage these systems in the face of fast-moving technology and a climate of increasing customer expectations. Design and management decisions require predictions of network performance; decisions based on poor predictions may adversely affect network customers' perception of the new technology. Analytical techniques, computer simulation, projections from existing experience, and experimentation, are methods that are used to evaluate and compare network designs and

protocols [FP01]. Independent of the prediction methodology, however, design and management decisions often must be made with incomplete knowledge of impending user demands and how the system will evolve.

In this chapter, we provide an overview of traffic and channel modeling for tele-communications networks, and queuing theory which allows us to present the work in a probabilistic framework.

3.2 Traffic and Channel Models

3.2.1 Traffic Models

Traffic modeling is a subject that has always generated considerable interest. Within the context of modeling and analysis of communications networks, the reason for this interest is clear. The performance of the network is highly dependent on the statistical features of the traffic presented to it [JR86].

A network that performs well for traffic that arrives according to a Poisson process may perform poorly with traffic that is bursty. A network that efficiently transports bulk data may be very inefficient with multi-media data [PF95]. The difficulty in accurately characterizing the traffic that will be presented to the network can be attributed to at least two factors [FM94]. Firstly, the demand on the network resources may be poorly understood. Secondly, the type of data on the network is constantly changing. Although voice, video and HTTP traffic accounted for only a modest level of the network traffic several years ago, they now dominate all other traffic types. Accurate performance modeling of a network has to presuppose knowledge of the application domain (e.g., telemetry, multi-media) that generated the network traffic.

Simple traffic consists of single arrivals of discrete entities (packets, cells, etc). It can be mathematically described as a point process, consisting of a sequence of arrival instants T_1, T_2, T_n,... measured from the origin 0; by convention, T_0 =0. There are two additional equivalent descriptions of point processes: counting processes and inter-arrival time processes [Cin75]. A counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t) = \max\{n : T_n \le t\}$ is the number of (traffic) arrivals in the interval (0,t). An inter-arrival time process is a non-negative random sequence $\{A_n\}_{n=1}^{\infty}$, where A_n=T_n-T_{n-1} is the length of the time interval separating the n^{th} arrival from the previous one. The equivalence of these descriptions follows from the equality of events

$$\{N(t) = n\} = \{T_n \le t < T_{n+1}\} = \{\sum_{k=1}^{n} A_k \le t < \sum_{k=1}^{n+1} A_k\}. \qquad (3\text{-}1)$$

Compound traffic consists of batch arrivals; that is, arrivals may consist of more than one unit at an arrival instant T_n .To fully describe compound traffic, one also needs to specify a non negative random sequence $\{B_n\}_{n=1}^{\infty}$, where B_n is the (random) number of units in the batch. At a higher level of abstraction, B_n may represent some general attributes of the n^{th} arrival, such as the amount of 'work' associated with the nth arrival or its mobility in a network. Such compound traffic processes are called marked point processes [FM94]. Discrete-time traffic processes correspond to the case when time is slotted. Mathematically, this means that the random variable A_n can assume only integer values, or equivalently, that the random variables N(t) are allowed to increase only at integer- valued time instants T_n .

Traffic processes are used to drive simulations in several ways, all of which use one or more pseudo-random number streams to generate sequences of random variables via appropriate transformations. To emphasize this point, we shall use the term 'randomly generated' to refer to such computer-generated random sequences. In the simplest case, a simulation only needs to randomly generate a sequence of inter-arrival times $\{A_n\}$.

In addition to arrival times and batch sizes, it is often useful (and sometimes essential) to incorporate the notion of workload into the traffic description. The workload is a general concept describing the amount of work $\{W_n\}$ brought to a system by the n^{th} arriving unit; it is usually assumed independent of inter-arrival times and batch sizes. A typical example is the sequence of service time requirements of arrivals at a queuing system; although in queuing, one usually refers to the arrival process alone as traffic. On the other hand, traffic reduces to workload description when inter-arrival times are deterministic. A case in point is compressed video, also known as VBR (variable bit rate) video, where coded frames (arrivals) have variable and random size (bit rate), and these must be delivered deterministically every 1/30 of a second or so for high-quality video. The workload consists of coded frame sizes (say, in bits), because frame size is roughly proportional to its transmission time (service requirement) [PF95].

The most commonly used stochastic model for packet arrivals is the Poisson model [FM94]. One of the reasons that the Poisson process has seen widespread use is that the memory-less property of exponential distributions makes analysis relatively simple since prior events do not affect the current probability of an event occurring. Additionally, since the combination of two or more Poisson processes results in another Poisson process, the analysis of multiple traffic sources is straight-forward. These compound

Poisson processes have been used to model batch arrivals where the inter-batch arrival times are independent and exponentially distributed.

It has long been recognized that packet arrivals in networks are not necessarily a Poisson process [JR86]. Recent studies have shown that wide-area network traffic is self-similar [LTWW94, PF95]. Self-similar traffic can be visually characterized by its scale-invariance. If packet arrivals per unit time is plotted in units of 10 seconds and compared to the same plot using units of 1 second, the burstiness of the inter-arrivals would look the same. Using a smaller time unit of 100 ms or 1 ms would result in plots that look the same as the larger time unit plot. In contrast, using smaller and smaller time units on plots of traffic that arrives according to a Poisson process would result in plots that at a larger time scale look relatively smooth and become more and more bursty as the time scale gets smaller ([LTWW94]).

It is proposed that the physical explanation for self-similar traffic is due to the superposition of many ON/OFF sources whose ON/OFF distributions have infinite variances. Several models that generate self-similar traffic have been proposed. A model based on doubly stochastic Poisson processes where the intensity of arrivals is modeled as a continuous stochastic process was proposed by [LS95]. The Random Midpoint Displacement (RMD) algorithm [LEWW95], which is scale-invariant, focuses on fast generation of self-similar traffic by recursively generating midpoint values (i.e., inter-arrival times). The RMD algorithm speeds up the process of choosing the values by picking the values independently at the time they are needed. Other self-similar traffic generation methods can be found in [Nor95].

By far the simplest way to generate self-similar traffic is to draw inter-arrival times from the Pareto distribution [PF95]. The Pareto distribution is a well-known heavy-tailed distribution with infinite variance, and is found to match very well with the actual data traffic measurements. The Pareto distribution was first used to describe the distribution of income among a population. The standard form for the two-parameter (location and shape) Pareto distribution function defined over the non negative real numbers can be written [FH99]as

$$p(t) = \frac{\alpha\beta^{\alpha}}{(t+\beta)^{\alpha+1}} \tag{3-2}$$

with an n_{th} moment $\quad E[T^n] = \frac{n!\,\beta^n}{(\alpha - n)!}$.

The Pareto distribution is heavy-tailed, which means that it is quite probable that a value far exceeding the mean will occur. The Pareto distribution has the characteristic that the mean and variance are infinite for $a \leq 1$, the mean is finite for $\alpha > 1$, and both the mean and variance are finite for $\alpha > 2$. Thus, no matter what the value of the parameter α is, a Pareto random variable cannot have all its moments and hence does not have a moment-generating function or Laplace transform. Therefore, since the Laplace transform plays such a vital role in the mathematical analysis of queuing systems, analysis possibilities become quite limited.

Another type of traffic model is the Markov-modulated model. In this type of model, different arrival probabilities are used for each of the k states in the Markov model. That is, each state, k, specifies a different process by which the probability of an arrival is determined. The amount of time spent in the state is 'modulated' by the underlying

Markov process. This type of model is also known as doubly stochastic in [FM94] and has been used to generate self-similar traffic in [SL95].

The ON/OFF model is widely-used to model bursty data such as voice traffic. Although ON/OFF models that can model speech events such as double-talk and mutual silence can be constructed, a simpler two-state Markov chain is often used to model voice traffic [Pru95, VZ95]. One state is the 'talk' state (T) and the other is the silent state (S).

Brady discovered that both talk and silence periods of digitized voice are exponentially distributed [Bra69]. The commonly accepted model for a speaker in a voice call is a continuous-time, discrete-state Markov chain. The holding time in each state is assumed to be exponentially distributed with mean $1/\beta$ and $1/\alpha$, respectively; hence the transitional rates from the ON to OFF state and from the OFF to ON state is k and s respectively (Figure 3-1). The commonly used values are $T_1 = 1/\beta = 650$ msec and $T_2 = 1/\alpha = 352$ msec. $1/\beta = 742.8$ msec and $1/\alpha = 435.7$ msec were also measured from actual telephone traffic with silence detectors of high sensitivity [MBK02]. A simple birth-death process can be used to model the situation of multiple calls in progress, with the state being the number of calls in talk spurt.

This model is a simplification of the full conversation model. After entering in the first state, emulation application starts to send packets into a network; and as long as it stays in that state, it continues to send equally spaced packets. Because of the discrete nature of this process, the next state is calculated after a packet is sent. This process can be described with a transition probability matrix given by $P = \begin{bmatrix} 1-k & k \\ s & 1-s \end{bmatrix}$.

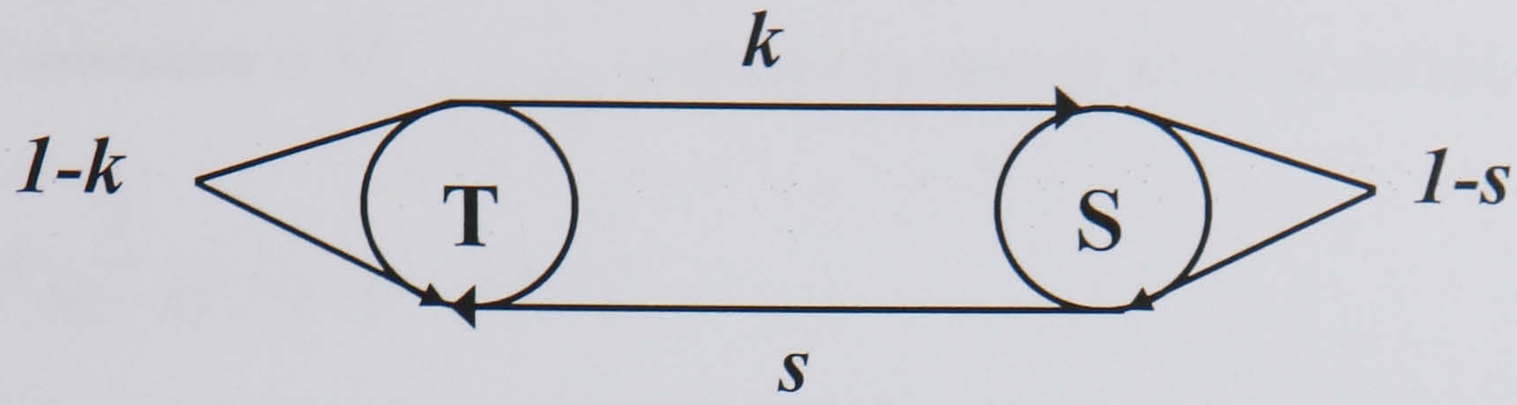


Figure 3-1: On/Off voice traffic model.

If the process is stationary, it can be showed (in the following proof) that $P_T = \frac{s}{k+s}$ is the probability of being in the talk spurt state, and $P_S = \frac{k}{k+s}$ is the probability of being in the silent state.

Proof: If the first packet of the talking interval has been generated, we can ask ourselves: "What is the probability that i packets will be generated in a sequence?"

According to Figure 3 1, this could be written as

$$P(i) = k(1-k)^{i-1}.$$

Now we can evaluate the mean value for this stochastic process as shown in equation (3-3). Notice here that the number of packets received in a sequence conforms to the geometric distribution

$$E[k] = \sum_{i=0}^{\infty} iP(i) = k\sum_{i=0}^{\infty} i(1-k)^{i-1} \; ; \qquad (3\text{-}3)$$

and thus

$$E[k] = k\frac{\partial}{\partial(1-k)}\sum_{i=0}^{\infty}(1-k)^{i} \; . \qquad (3\text{-}4)$$

It can easily be seen that equation (3-4) represents geometrical series. Therefore, the result of summation is $1/(1-(1-k))$, and hence equation (3-4) can be verified as

$$E[k] = k\frac{\partial}{\partial(1-k)}[1/(1-(1-k))] = \frac{1}{k}. \qquad (3\text{-}5)$$

Since packets in the talk interval are sent equally spaced by Δ ms, we can write:

$$T_1 = \Delta\frac{1}{k} \quad \text{and} \quad T_2 = \Delta\frac{1}{s},$$

where T_1 and T_2 are mean times spent in state T and S, respectively. With the stationary assumption, P_S and P_T can be given by $\frac{T_2}{T_1+T_2}$ and $\frac{T_1}{T_1+T_2}$ respectively.

In the case of multiple voice sources it is necessary to introduce a general model derived from the basic two state Markov model. Thus, it is observed that the steady-state arrival process is a binomial process. This is clearly due to the fact that each speaker's behavior is an independent Bernoulli trial. The question is 'How to express the probability that in the k_{th} time interval (interval duration is fixed to Δ ms) n of m sources are active?' For steady-state, or in other words, when $k \to \infty$ one can write

$$P_n = \binom{m}{n}\left(\frac{s}{k+s}\right)^n\left(\frac{k}{k+s}\right)^{(m-n)}. \qquad (3\text{-}6)$$

The expected value of n (the average number of voice sources which are in talk spurt) can be written as $E(n) = mP_T$.

3.2.2 Channel Models

A common figure of merit used in digital links is the bit-error-rate (BER): the probability that a bit is received in error. The BER for a digital link is analogous to

signal-to-noise ratio (SNR) for analog links [PB86]. Two types of BERs are commonly used in modeling channels. A static BER remains constant during the entire time the model is being used. A dynamic BER can change based on some parameter such as elapsed time or the number of bits transmitted. Static BER models assume that bit errors are statistically independent. It is well-known that errors in IP networks tend to occur in 'bursts' and therefore cannot be accurately modeled using the assumption of independent errors [DMM88].

The classic dynamic BER model for digital channels is the Gilbert model [Gil60]. The Gilbert model is based on a two-state Markov chain shown in Figure 3-2. In the G or 'good' state, no bit errors occur. In the B or 'bad' state, errors occur with probability (1–h) where h is the probability of no bit error. A G-to-B state transition occurs with probability p; a B-to-G transition occurs with probability q. The model remains in state G with probability $1-p$ and remains in state B with probability $1-q$. This model has been shown to reproduce errors that occur in an IP network more accurately than a static BER model.

The probability of a state transition in the Gilbert model is evaluated upon the presentation of a bit to the channel. That is, state transitions are evaluated on a bit-by-bit basis. This type of model can be termed a 'transmission modulated model'. In the context of simulation, this may require an inordinate amount of computation. A common technique to reduce this computational burden is to model the number of bits between state transitions as a geometrically distributed random variable. Therefore, rather than evaluating each bit for a possible state transition, a single calculation gives the number of bits between state transitions. Consider a 1 Mbps wireless channel with an average BER

of10^{-6}. To observe a single error, an average of 10^{6} bits must be transmitted while the channel is in the bad state. Of course, if the channel is in a good state, no errors occur.

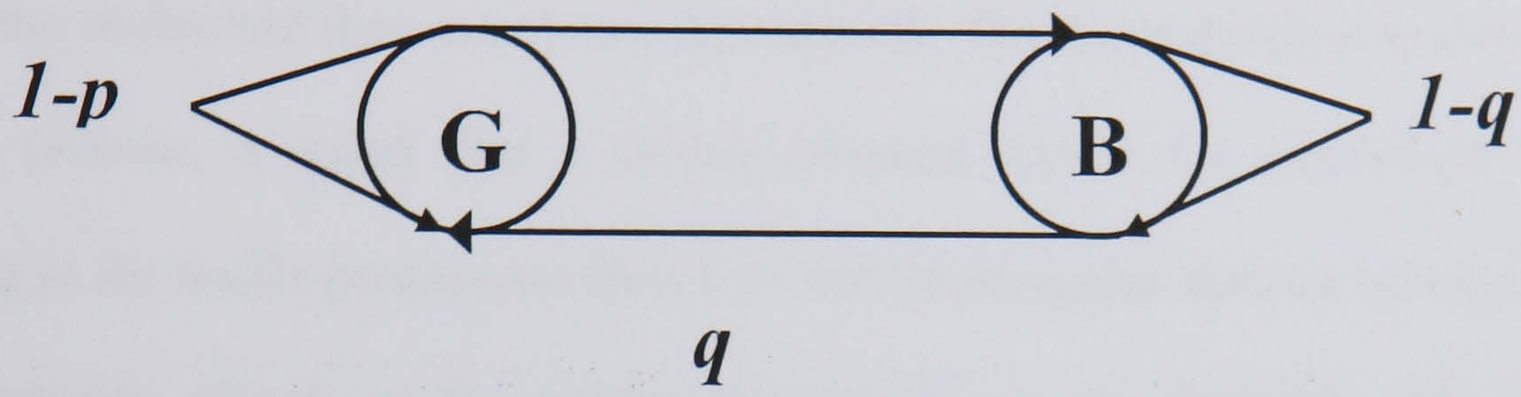


Figure 3-2: Gilbert Model Transition Diagram.

Furthermore, a transmission modulated model makes the impractical assumption that the state of the channel does not change when there are no bits in the channel. An alternative to a transmission modulated model is a time modulated model. In this type of model, state transitions occur based on elapsed time rather than the number of bits transmitted. Using the two-state Gilbert model as an example, the time spent in the good or bad state is modeled as an exponentially distributed random variable with different means [YJ02]. It can be seen the probability that ***n*** consecutive packets are lost equals $q(1-q)^{n-1}$ and thus, the residence time for state B is geometrically distributed. The loss rate according to the Gilbert model is

$$P = \frac{p}{p+q}. \qquad (3\text{-}7)$$

This significantly reduced the computational burden and appeals to the impression that the state of the channel does indeed change even though no bits are being transmitted.

Research that has used this approach to channel modeling includes [BBKT97].

3.3 Queuing Theory

A queuing network is a collection of two or more single queues or 'nodes' where customers receive service. Customers arriving at the network request service at one or more of the nodes and then may leave the network. The basic queuing system comprises a traffic process, a queue and a service element. Tasks (or customers) that arrive according to the traffic process are forced to wait in the queue and are serviced according to the demands placed on the service element. If we can describe each stage in the queuing system adequately, queuing theory provides results that will allow us to gain insight into the dynamic nature of such systems [Kle76]. The disadvantage with a queuing theory approach is that detailed analysis of queuing problems is intractable for all but relatively simple models.

Classification is especially important in queuing networks. Many classes of networks have no known closed-form solutions. Other networks have state spaces that are so large that certain analysis techniques, while theoretically possible, become intractable. For these cases, approximations (or perhaps simulation) may be appropriate. The following sections introduce terms and concepts used to classify queuing networks.

Our analysis will be based on finding a task value of a queuing system for various queuing disciplines. In particular, we are interested in the waiting time of the task; and even the system time, which is the sum of the waiting time and processing time. A derivation of expressions for the waiting time probability distribution function (PDF) for the M/M/1 FCFS queue was first derived by Erlang in 1909, and derived for the M/G/1 queue by Pollaczek and Khintchine [BGMT98].

In order to name the different kinds of queuing systems to be distinguished, a rather simple shorthand notation is used for describing queues. This involves a three-component description, A/B/m, which denotes an m-server queuing system where A and B 'describe' the inter-arrival time distribution and service time distribution respectively. A and B take on values from the following set of symbols, which are meant to remind the reader which distributions they refer to: M and G imply exponential and general PDFs, respectively [Kle76]. According to the traffic model in the packet network, one of the interesting systems we consider in this chapter is the G/M/1 queue in which we have arbitrary inter-arrival times, exponential service times, PDFs and a single server.

3.3.1 The G/M/l Queue System

The G/M/1 system is in fact the 'dual' of the M/G/I system. Surprisingly, G/M/I yields to analysis more easily than M/G/l and so we can quote distributions directly. The system, of course, corresponds to the case of an arbitrary inter-arrival time whose PDF is given by A(t), with pdf a(r) and Laplace transform of which is denoted by ***A*(s)***, (Laplace-Stieltjes transform of the PDF A(t)); and service times distributed exponentially with mean $\frac{1}{\mu}$. All the results are expressed in terms of a root σ that is the unique root in the range $0 \le \sigma < 1$ of the functional equation

$$\sigma = A^*(\mu - \mu\sigma). \qquad (3\text{-}8)$$

Once σ is evaluated, the following results are immediately available. The distribution for the number of customers found in the system by a new arrival is given by

$$r_k = (1-\sigma)\sigma^k \qquad k = 0,1,2,... \qquad (3\text{-}9)$$

the PDF for waiting time is given by

$$W(y) = 1 - \sigma e^{-\mu(1-\sigma)y} \qquad y \geq 0 \ ; \tag{3-10}$$

and the mean waiting time is given by

$$W = \frac{\sigma}{\mu(1-\sigma)} \ . \tag{3-11}$$

It is remarkable that the waiting times are exponentially distributed, independent of the form of the inter-arrival time distribution [BGMT98].

3.4 Summary

This chapter highlighted different IP network channel and traffic models. Basic aspects of the Poisson process and self-similar traffic models were presented in Section 3.2.1. Also, in this section, different kinds of stochastic models for data and voice traffics, involving the concept of self-similarity and the long-range dependence nature of the network traffic, have been reviewed. The most common and classic dynamic BER model for digital communication; as well as the Gilbert channel model, which is based on a two-state Markov chain, were discussed in Section 3.2.2.

The basic queuing system concept and a simple shorthand notation, which is used for describing queues, have been explained in Section 3.3. At the end of this chapter (Section 3.3.1) we expressed in details the G/M/I queue system which can be mostly matched to our problem.

Chapter 4- Voice Codec-base Redundancy Schemes

4.1 Introduction

The quality of service is a major factor in preventing the Internet telephony from competing with the traditional circuit-switched telephone. The current Internet is designed for traditional digital data transmission; it monitors its concerns about the overall transmission throughput and relative reliability by employing the BE service. However, the TCP/IP protocol and the BE service model are not suitable for real-time streams, since they cannot provide any bandwidth or delay guarantees. The burstiness is an innate property of the TCP traffic. In order to improve the bandwidth utilization, most routers use large size buffers to absorb the bursty TCP traffic. This poses a problem for real-time packets since they will miss their deadlines, during congestion, because of being buffered for extended periods of time [Tho96, HNA00].

As we discussed in Chapter 2, RTP headers, such as time stamp and sequence number, are then added to form an RTP audio packet. The packet is then sent as a UDP packet to the receiver through the Internet. If a packet gets lost or misses the deadline, the receiver will generate a predicted packet based on the neighboring packets. Retransmission is generally not a proper approach, for it implies that the deadline has been missed in most cases. To provide a delay bound for the real-time packets, the packet is isolated from the TCP traffic. The real-time services require limited buffer size in the routers so that the packet can have a total bounded delay.

Voice traffic can tolerate some amount of packet loss; a packet loss rate greater than 5% is considered harmful to the voice quality [PHH98]. The amount of packet loss rate that can be tolerated depends on the nature of the encoding algorithm and on the sampling rate of the voice stream. The length of a phoneme is typically from 80 to 100ms [Lu00], and a loss greater than the length of a phoneme can change the meaning of a word. Changing the IP infrastructure to support session's with guaranteed bandwidth would allow for effective transport of voice streams. Changing the Internet infrastructure is a difficult proposition; hence the interest has been in application-level techniques for compensating of packet delay jitter and loss.

Algorithms incorporating FEC methods have been developed to compensate for packet loss [BPT99]. However, existing methods can be shown to have shortcomings in their response to burst losses of packets and may exhibit unstable behavior. We propose a method that invokes the VCBR control algorithms (redundancy sending), and a new error correction method which uses VCBR with the backup channel (VCBRBC); and employ a source coder to reduce the overall computational complexity as compared to that in the method proposed by Bolot *et al* [BG96].

In addition we will show that the bandwidth efficiency increases by at least 13 percent and tolerable packet loss can be increased to less than 10 percent. The total required capacity, however, only increases by less than codec bit rate.

In this chapter, our proposed codec-specific VCBR and VCBRBC are introduced; and the respective channel is modeled by the Gilbert loss model. At the end, we evaluate the performance of the proposed error correction technique using the Network Simulator NS2 [FV05].

4.2 VCBR and VCBRBC Algorithm

4.2.1 VCBR Algorithm

The FEC methods can be used for voice applications in two ways. The first is to employ a general purpose FEC technique that is applied to the voice packets. In this method a number of voice packets may be protected by a FEC code that is sent as a separate packet or block of packets. This introduces a delay dependent on the length of the respective block of voice packets. The second is to use the FEC as a part of the voice codec.

The latter produces two versions of the voice code; regular code, and a more compressed code that can be used in case of errors in the regular code. In the codec-specific FEC, there is a slight degradation in speech quality at the receiver, but not as significant degradation as if the packet were completely lost [BPT99]. If only one source coder for both the main and the redundant payload is used, the overall computational complexity for reconstruction can be reduced. Moreover, the packet loss is mitigated without introducing more congestion, and hence, a more scalable and effective approach is offered than successively adding redundancy to a constant bit-rate source.

Voice codec-base redundancy (VCBR) schemes piggy-back information about the present period, with later packets. This is shown in Figure 4-1 as the secondary encoding in a packet (with reconstruction occurring at the receiver). Although this technique uses lower bit rate codecs, studies show that the quality of the reconstructed stream is quite good and this scheme increases intelligibility [PHH98-HNA00]. If packet n carries a redundant encoding of packet n-i, and if packet n-i is lost, the application waits for packet

n to recover the lost packet. Thus, a single lost packet can be recovered with i packets worth of delay.

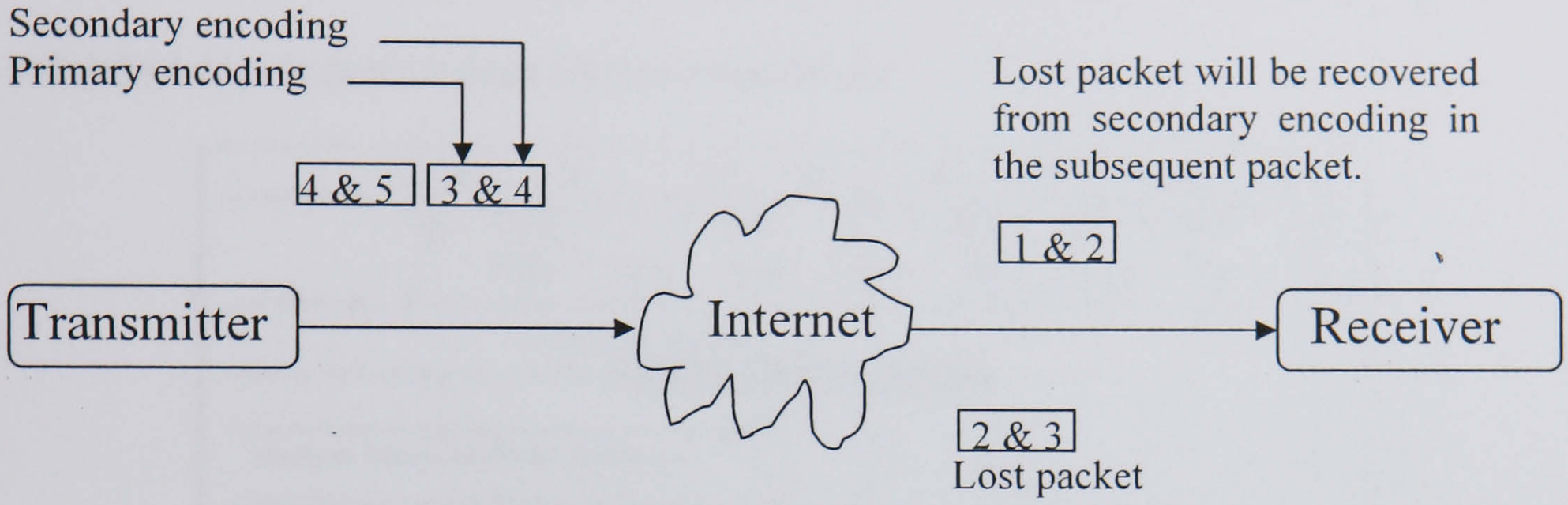


Figure 4-1: Voice odec -base redundancy scheme.

Two VCBR schemes will be employed to evaluate the speech-property base Firstly, the two frames of the packet *n* are piggy-backed on the packet *(n+2)* (to further mitigate the effect of packet burst loss we do not piggy-back the two frames of the packet *n* on the packet *(n+1)*). This scheme has a redundancy overhead of *100%*; but it is simple for the sender to construct and for the receiver to use it for recovery of the lost frames. In the second VCBR scheme, the four frames of packet *(n)* and *(n+1)* are Xored and the result is piggy-backed on the packet *(n+2)*. If the packet *(n+2)* and one of the packets *(n)* or *(n+1)* arrive at the receiver, a lost packet can be recovered. The later has a redundancy overhead of *50%*.

As in Section 2-5, the compression algorithms work by analyzing a block of PCM samples delivered by the Voice codec. These blocks vary in length depending on the coder. For example, the basic block size used by a G.729 algorithm is 10 ms whereas the basic block size used by the G.723.1 algorithms is 30ms (Table 2-3). An example of how

a G.729 compression and packetization system work is shown in Figure 4-2. As it can be seen a block of 3 frames takes more than 30 ms because of the codec delay. We use G.729 as the codec and assume two frames in a packet (20 ms), and 2 frames of redundant data to protect these frames respectively.

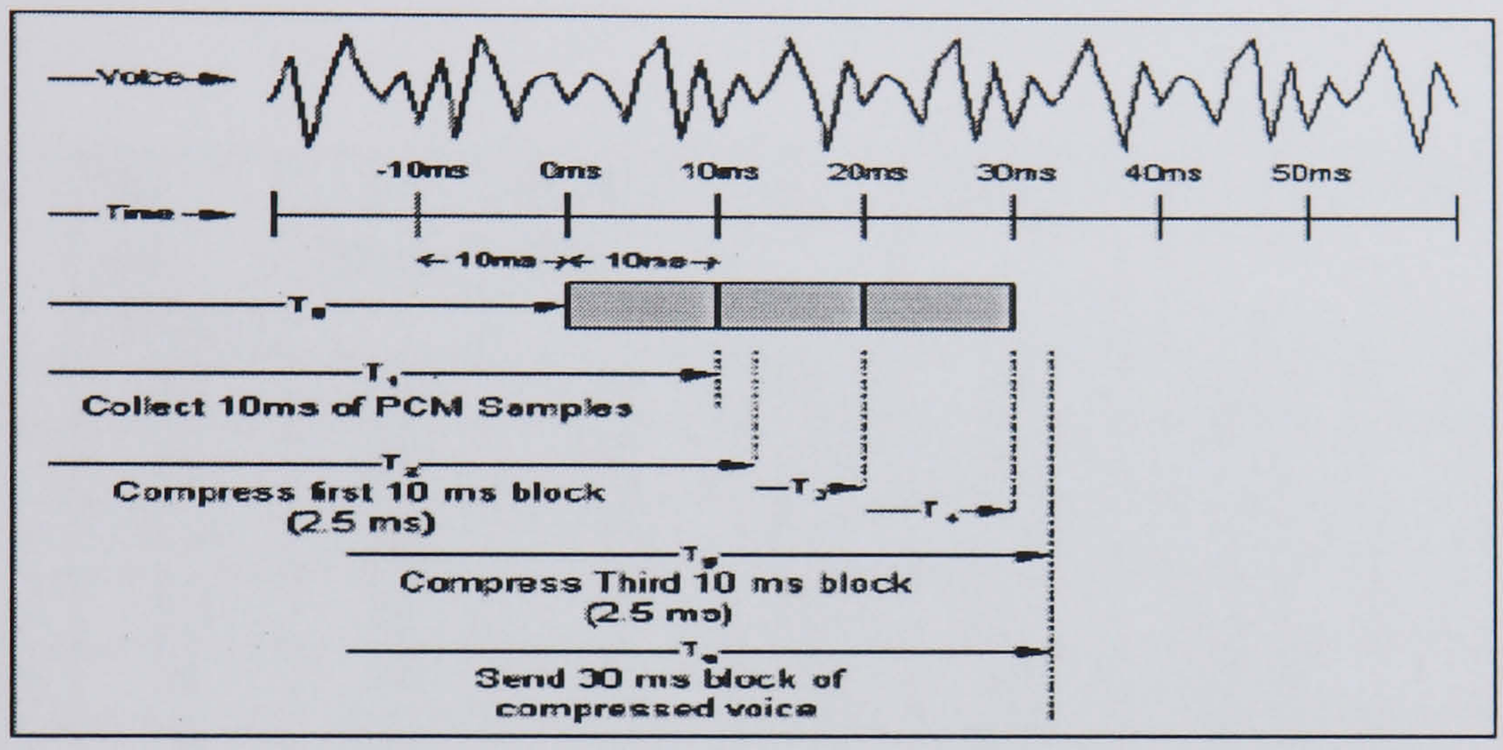


Figure 4-2: Voice compression and packetization.

If we have k *coded* frames with frame size t millisecond and frame length l bytes within a packet which contains UDP/RTP/IP, the bandwidth efficiency $\mathcal{B}_{eff}$ and required capacity $\mathcal{R}_{cap}$ for transport can be defined as

$$\%\mathcal{B}_{eff} = 100\frac{kl}{kl+40} \tag{4-1}$$

$$\mathcal{R}_{cap} = 7.8125\frac{kl+40}{t} \quad \text{Kbit/sec}. \tag{4-2}$$

If only one source coder for both the main and the redundant payload is employed, it can see some reduction in the overall computational complexity. Moreover, in the cases where the primary and the redundant data of a packet are coded using different audio encoding schemes, and 'piggy-backing' on the following packets is employed; all decoders suffer loss of synchronization, and deliver decoded speech signals with poor

quality in the event of a loss of an important frame. It can be seen from table 4-1 that using the first scheme for *low bit rate codec,* the bandwidth efficiency increases by 17 percent, and only one packet is lost when packets *n* and *(n+2)* have been lost. In other words, the level of loss tolerance by the network has been increased by more than 50 percent.

Codec	Bit rate (Kbps)	*k l* (Byte)	Frame size	R_{cap}	$\%B_{eff}$	R_{cap} VCBR	$\%B_{eff}$ VCBR
G.711	64	160	20	78.1	80	140.6	88.8
G.726	32	80	20	46.8	66.6	78.1	80
G.723.1	6.4	24	30	16.7	37.5	22.9	54.5
G.723.1	5.3	20	30	15.6	33.3	20.8	50
G.729A	8	10	10	39	20	46.8	33.3
G.729A	8	20	20	23.4	33.3	31.2	50

Table 4-1: Some important codecs and their details.

4.2.2 VCBR using Backup Channel (VCBRBC)

Redundant information is transmitted along with the original information in VCBR, so that the lost original data can be recovered at least in part from the redundant information. Although sending additional redundancy increases the probability of recovering lost packets, it also increases the required capacity and thus the loss rate of the audio stream in the event of congestion. The results in Section 4.2.1 show that the bandwidth efficiency and the loss tolerance can improve by 17% and 50%, respectively, through increasing the required capacity for voice stream by less than coding rate.

In many practical cases, consecutive packet losses are correlated, due to the way packets are dropped. A packet loss can be followed by a burst of loss, which significantly decreases the efficiency of VCBR schemes. On the other hand, VCBR methods have

the network [LSG01]. Moreover, the end-to-end delay is increased in most cases by applying such methods. Therefore a more scalable and effective approach than successively adding redundancy to a constant bit-rate source, is introduced to avoid worsening the network congestion. Moreover, the packet loss is mitigated without introducing more congestion. If the packet loss is caused by congestion or bursty phenomena, the redundant information which is piggy-backed with the main packet could worsen the loss in the network. the feasibility of improving the performance of end-to-end data transfers between different sites through path switching.

On the other hand, the performance of end-to-end data transfers between different sites through path switching [TXXF04] and delay-loss tradeoff for real time voice communication over the Internet are improved by taking advantage of the highly uncorrelated delay variation over multiple independent network paths [AS04, TXEG05]. Packet loss in delay-sensitive applications such as interactive VoIP is a result of not only packet erasure, but also delay jitter. By sending the redundant information through an uncorrelated network path, rather than transmitting together with the main voice packet (Figure 4-3), the receiver can often retrieve those packets which experience erasure or excessive delay through the main path. Latency can also be reduced in this approach by playing out the voice description through the path with lower delay, without any excessive computational algorithm; compared with multi-stream voice transmission over multiple paths.

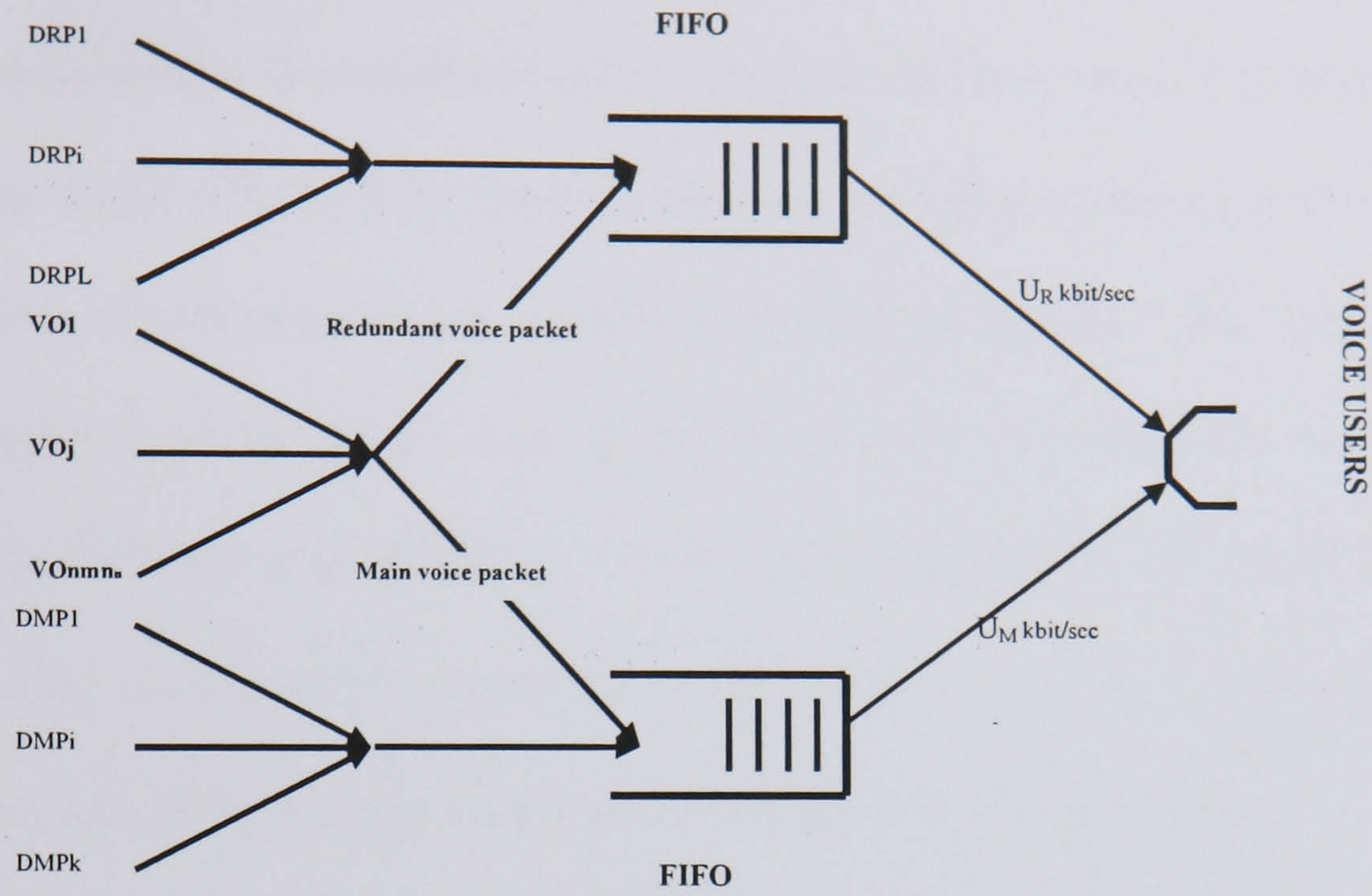


Figure 4-3: VCBRBC schematic diagram for transmitting Data over Redundant Path (DRP) with U_R received bit rate, and Data over Main Path (DMP) with U_M received bit rate.

The efficiency of the backup channel depends on the statistical correlation between channels. It also depends on the availability of another channel option which has a mean delay not much higher than that of the main path [TXXF04]. This can be obtained in practice by sending streams over networks with close geographical routes which are serviced by different Internet Service Providers.

4.3 Gilbert Loss Model for VCBR and VCBRBC Techniques

The Gilbert model has been in wide spread use in the literature of transport and error correcting in the computer communication systems. The Gilbert model is often employed to capture temporally correlated loss in the Internet [YJ02]. The Gilbert loss model

(GLM) has also been used to assess the VCBR techniques within the context of real-time media application as well as the reliability of data transfer over multicast IP networks.

In the two-state GLM (Figure 3-2), the time spent in the good or bad state is modeled as an exponentially distributed random variable with different means. This significantly reduces the complexity of the model, and appeals to the impression that the state of the channel does indeed change even though no bits are being transmitted. Due to this model, the probability that n consecutive packets are lost equals $q(1-q)^{n-1}$, and thus, the residence time for state B or G is geometrically distributed. If we assume that:

$X|B$: *is the number of consecutive lost packets given state B,*

The conditional probability density function for state B can be written as

$$P(X|B) = q(1-q)^{X-1}. \tag{4-3}$$

The conditional expected value of X, $E[X|B]$ can then be expressed as

$$E[X|B] = \sum_{X=0}^{\infty} XP(X|B) = q\sum_{X=0}^{\infty} X(1-q)^{X-1}. \tag{4-4}$$

Hence, $E[X|B]$ and $E[Y|G]$ can be given, it has been calculated in equations (3-4) and (3-5), as

$$E[X|B] = \frac{1}{q} \text{ and } E[Y|G] = \frac{1}{p},$$

where $Y|G$ *is the number of consecutive received packets given state G.*

So, in the stationary conditions, the loss rate P_l according to the GLM can be written as

$$P_l = \frac{1/q}{1/p+1/q} = \frac{p}{p+q} \tag{4-5}$$

The level of achievable improvement through adding redundancy to the transmitted packets will determine the amount of required redundant information at each instant. Consider now the case when only the *n*th packet includes redundant information about the (*n*-1)th packet. A packet is lost only if it cannot be reconstructed using the redundant information; that is, when the packet is lost and the following packet is also lost. It is then straight-forward to show that the loss rate after reconstruction is $\frac{p(1-q)}{p+q}$.

Similar analysis can be used for examining cases with two, three or more pieces of redundant information. The results are summarized in table 4-2. In this table we have used a similar notation as in [BPT99]. In the column 'Redundancy' the notation '-1,2' for example, means that redundant information about the (*n-1*)th and (*n-2*)th packets are sent in the *n*th packet.

However, using the same model for the VCBRBC method and equation (4-5), it can be seen that the loss rate for the overall system will be $P_{l1}.P_{l2}$, in which $P_{l1}=\frac{p_1}{p_1+q_1}$ and $P_{l2}=\frac{p_2}{p_2+q_2}$ are the loss rates for path 1 and path 2, respectively. With a simple comparison, the loss rate of the VCBRBC is revealed to be less than that of the other VCBR schemes in table 4-2.

Redundancy	Loss rate after reconstruction
None	$p/(p+q)$
-1	$(p(1-q))/(p+q)$
-2	$(p^2q + p(1-q)^2)/(p+q)$
-1 -2	$p(1-q)^2/(p+q)$
-1 -3	$(p(1-q)(pq+1-2q+q^2))/(p+q)$
-1 -2 -3	$(p(1-q)^3)/(p+q)$
Two independent Paths,(VCBRBC)	$\frac{p_1 \cdot p_2}{(p_1+q_1).(p_2+q_2)}$

Table 4-2: Loss rate regarding the Gilbert loss model

4.4 Numerical Results

The efficiency and performance of the VCBR and VCBRBC will be evaluated in this section.It is shown through simulation results that using our proposed VCBR and backup channel, one can considerably improve the packet loss and delay performance of the VoIP networks. Network simulator NS2 will be used for the simulation.

NS2 is an object oriented simulator, written in C++, with an OTcl interpreter as a front end. The simulator supports a class hierarchy in C++, and a similar class hierarchy within the OTcl interpreter. The two hierarchies are closely related to each other; from the user's perspective, there is a one-to-one correspondence between a class in the interpreted hierarchy and one in the compiled hierarchy. Users create new simulator objects through the interpreter; these objects are instantiated within the interpreter, and are closely mirrored by a corresponding object in the compiled hierarchy [FV05].

4.4.1 VCBR Evaluation

We begin by considering, for a voice codec-base redundancy evaluation, a simple and single congested link between edge and core routers, which is shared by all flows. The linear topology (Figure 4-4) consists of core links of 1 Mbps capacity which act as the bottleneck link [FNYF03]. The topology is constructed to provide four main 300 Kilobit (Kb) per second (sec) links for voice traffic (nodes 0,1,2,3 which terminate at node 4 via shared link between nodes 7 and 8). Each of these four links may convey the traffic of more than one customer, depending on the capacity of the link and the codec type (Table 4-1). Node 5 is used for file transfer protocol (FTP) traffic, and terminates at node 6, with a single shared 1 Mbps bottleneck link in the middle (between nodes 7 and 8).

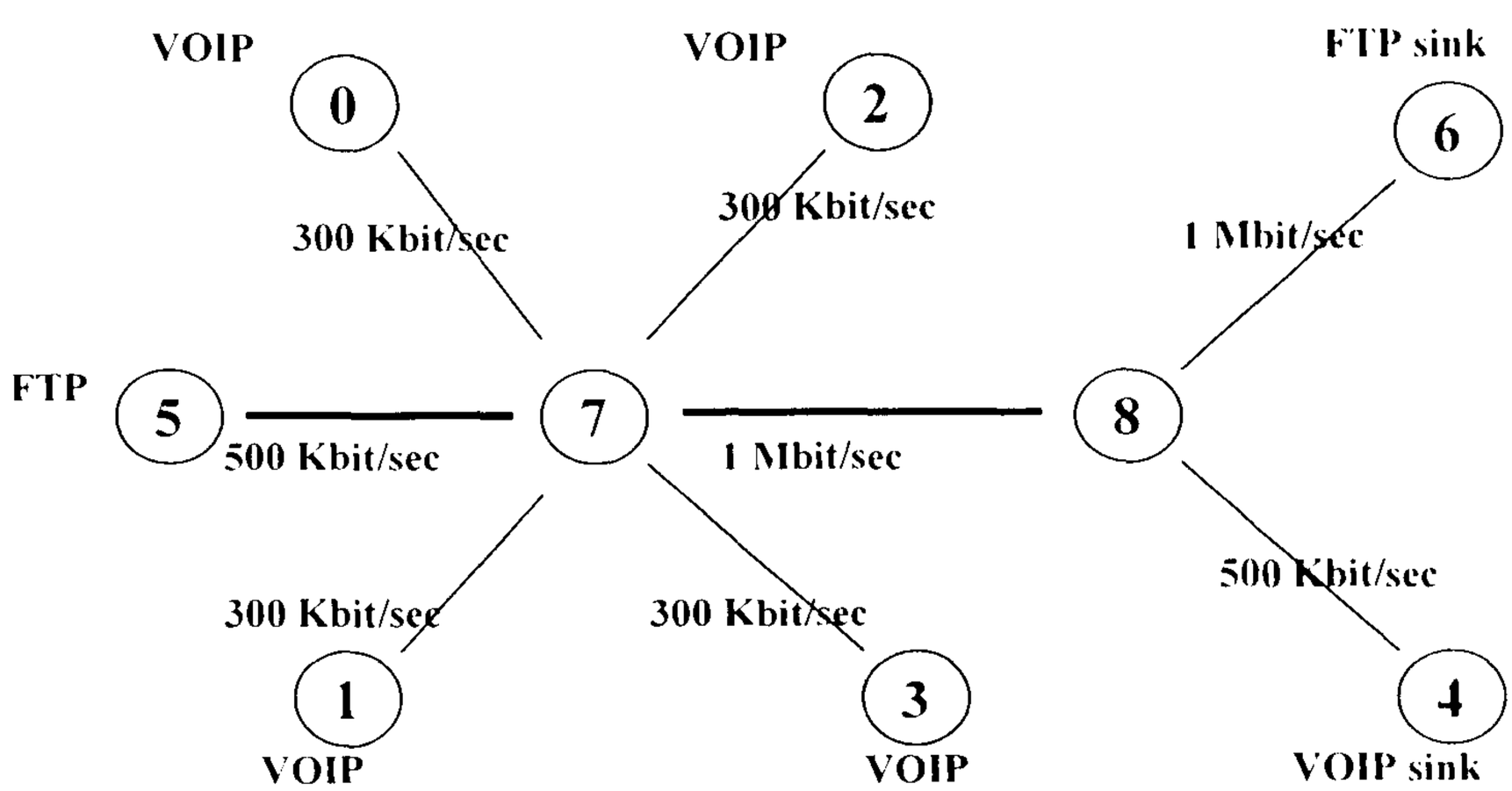


Figure 4-4: The network structure used in simulation.

We choose an FTP source over a TCP connection between nodes 5 and 6 at 500 Kb per sec and packet size 1000 bytes, which start at 0.1 secondin the simulation. An exponential traffic generator is employed that generates traffic according to an

exponential On/Off distribution over an RTP connection for voice links. Packets are sent at a fixed rate during on periods, and no packets are sent during off intervals. Both on and off periods are taken from an exponential distribution, with the same parameters as follows [MBK02]:

- The constant size of the packets is 80 bytes.
- The average 'on' ('off') time for the generator is 500 msec.
- The sending rate during 'on' ('off') times is 300 Kbit/sec.

20-byte voice frames, encapsulated in an RTP with 12-byte header, 8-byte UDP; and 20-byte IP in sequence, are generated every 20 msec. The total capacity required for this scenario is 23.43 Kbit per second with 480 -bit packet size (equation 4-2). Having 300 Kbit per second capacity allows 13 voice sources to be delivered. However, voice sources with voice activity detection technology do not send packets at every voice frame interval. The exponential source can simulate this situation; and on this condition, 22 voice sources were delivered in the simulation. Also there is a drop-tail (FIFO) queue management with a maximum buffer size of 15 in the link between nodes 7 and 8. Figure 4-5 shows the network simulator output in which the transmission of packets and status of the queue can be seen.

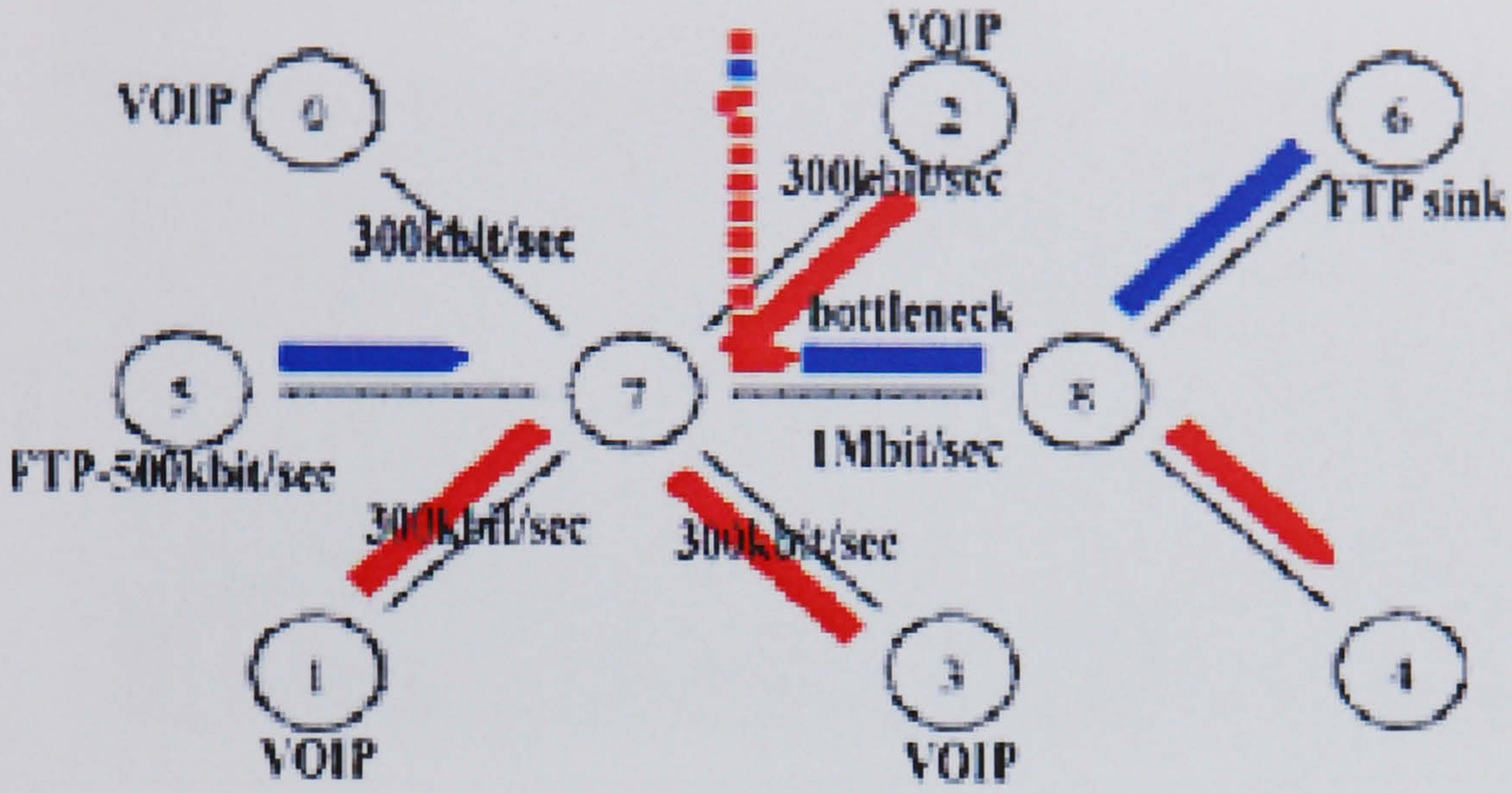


Figure 4-5: The simulation output of the network simulator.

As mentioned earlier and can be seen in Table 4-1, employing an VCBR scheme implies 80-byte packet size, and require 31.2 Kbit per second capacity. For 300 Kbit per second capacity and exploiting voice activity detection technology, 17 voice sources can be delivered. The simulation was run only on the best-effort condition. Figure 4-6 depicts the number of transmitted packets over time; and the number of lost packets over time is shown in Figure 47.

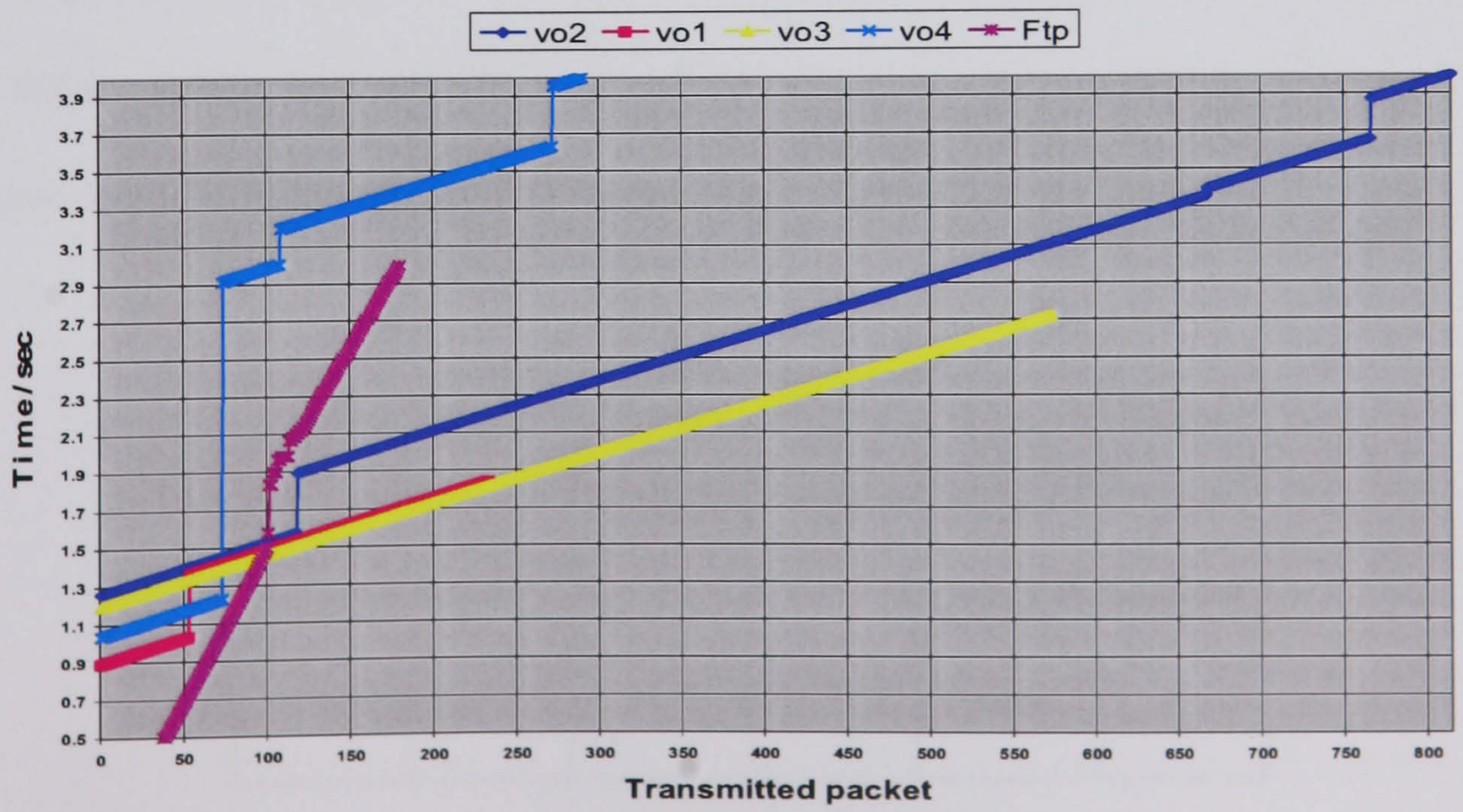


Figure 4-6: Number of transmitted packets for all traffic versus time.

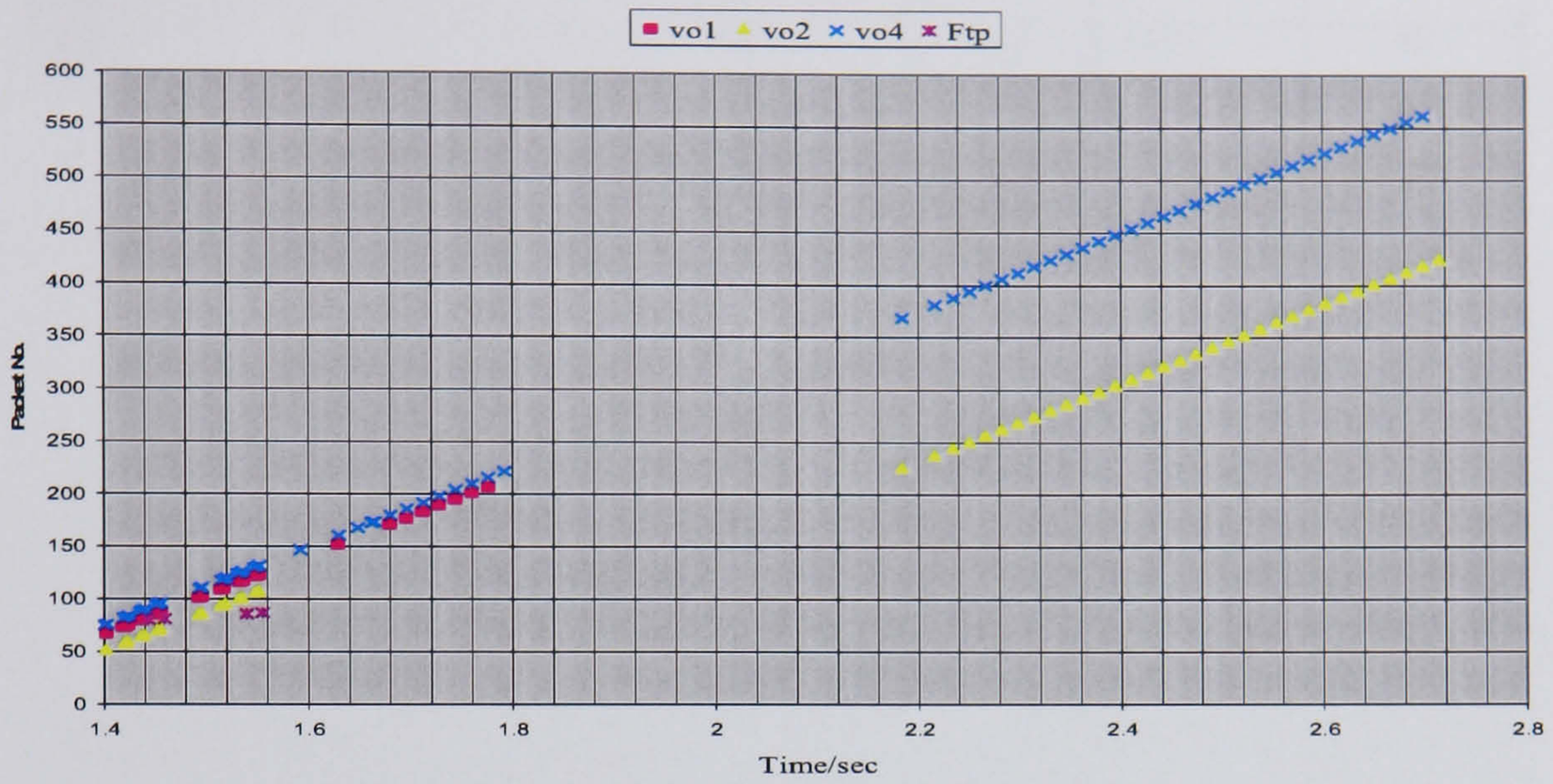


Figure 4-7: Number of dropped packets for all traffic versus time.

Tables 4-3 and 4-4 show the number of lost packets (out of the whole traffic packets), with and without our approach, respectively, in which a redundant packet is created and the two frames of the packet *(n)* are piggy-backed on the packet *(n+2)*. The performance improvement of the new VCBR approach can be seen by the massive decrease in the number of packets lost (compare Table 4-3 with Table 4-4). So in this case we can increase the rate of the codec to improve the quality of the voice, or raise the number of voice users within the network to exploit the network resources more efficiently. However, the lost FTP packets will be recovered in the TCP procedure.

V_{o1} time	V_{o1} No.	V_{o2} time	V_{o2} No.	V_{o4} time	V_{o4} No.
1.418	73	1.513	93	1.397	74
				1.5306	124
				1.5466	130

Table 4-3: Packets lost using VCBR. $V_{oi}time$:Time in which the packets are lost in the i^{th} voice traffic; $V_{oi}No$:packet number which is lost in the i^{th} voice traffic.

$V_{o1}time$	$V_{o1}No$	$V_{o2}time$	$V_{o2}No$	$V_{o4}time$	$V_{o4}No$	Ftp_time	Ftp_No
1.3809	59	1.3804	43	1.3973	74	1.4	80
1.3995	66	1.3991	50	1.4	75	1.4594	81
1.402259	67	1.401768	51	1.402667	76	1.5394	86
1.418259	73	1.420435	58	1.418667	82	1.5554	87
1.4209	74	1.4231	59	1.4213	83		
1.4235	75	1.4364	64	1.424	84		
1.4369	80	1.4391	65	1.4346	88		
1.4395	81	1.4524	70	1.4373	89		
1.452	86	1.455	71	1.44	90		
1.455	87	1.492	85	1.450	94		
1.492	101	1.495	86	1.453	95		
1.495	102	1.513	93	1.456	96		
1.514259	109	1.516435	94	1.493333	110		
1.516	110	1.519	95	1.496	111		
1.532	116	1.532	100	1.514	118		
1.535	117	1.535	101	1.517	119		
1.548	122	1.548	106	1.52	120		
1.551	123	1.551	107	1.530	124		
1.628	152	2.183	225	1.533	125		
1.679	171	2.215	237	1.536	126		
1.695	177	2.234	244	1.546	130		
1.711	183	2.250	250	1.549	131		
1.727	189	2.266	256	1.552	132		
1.743	195	2.282	262	1.592	147		
1.759	201	2.298	268	1.629	161		
1.775	207	2.314	274	1.648	168		
		2.330	280	1.661	173		
		2.346	286	1.664	174		
		2.362	292	1.68	180		
		2.378	298	1.696	186		
		2.394	304	1.712	192		
		2.410	310	1.728	198		
		2.426	316	1.744	204		
		2.442	322	1.76	210		

		2.458	328	1.776	216		
		2.474	334	1.792	222		
		2.490	340	2.184	369		
		2.506	346	2.216	381		
		2.522	352	2.234	388		
		2.538	358	2.250	394		
		2.554	364	2.266	400		
		2.570	370	2.283	406		
		2.586	376	2.298	412		
		2.602	382	2.314	418		
		2.618	388	2.330	424		
		2.634	394	2.346	430		
		2.650	400	2.362	436		
		2.666	406	2.378	442		
		2.682	412	2.394	448		
		2.698	418	2.410	454		
		2.714	424	2.426	460		
				2.442	466		
				2.458	472		
				2.474	478		
				2.490	484		
				2.506	490		
				2.522	496		
				2.538	502		
				2.554	508		
				2.570	514		
				2.586	520		
				2.602	526		
				2.618	532		
				2.634	538		

Table 4-4: Packet lost without using VCBR. $V_{oi}time$ Time in which the packets are lost in the ith voice traffic; $V_{oi}No$: Packet number which is lost in the ith voice traffic.

4.4.2 VCBRBC Evaluation

A linear topology has been considered for paths 1, 2, and 3, depicted in Figure 4-9. Each path has two congested links between edge and core routers, namely node zero and four, which are shared by data flows. The capacitys of the two congested links are 1 and 2 Mbps, 1.2 and 1.9Mbps, and 1 and 2.2 Mbps over paths 1, 2, and 3, respectively. The data traffic is produced by two FTP and two Pareto traffic generators which pass through the congested links within each path. Also the voice traffic aggregates consist of packets generated by sources which use the same codec schemes.

As in Section 3.2.1 one well known heavy-tailed distribution with infinite variance is the Pareto distribution. This distribution has been shown to match very well with the actual data traffic measurements with a typical α=1.35, for HTTP (hypertext transfer protocol) or web applications in a WAN (wide area network) [Pru95]. The probability density function of a Pareto-distributed variable T is $p_T(t)=\frac{\alpha}{(t+1)^{\alpha+1}}$, where α denotes the 'shape' parameter.

On the other hand, the commonly accepted model for speech during a voice call is a continuous-time, discrete-state Markov chain. The holding time in each state (talk and silence) is assumed to be exponentially distributed with mean $1/\beta$ and $1/\alpha$; with commonly used values of 650 milliseconds and 352 milliseconds respectively [MBK02].

For each path we have chosen an FTP source over the TCP connection between nodes 2 and 4, and between nodes 6 and 4 (Figure 4-10). The On/Off Pareto distribution is used

between nodes 3 and 4, and between nodes 7 and 4, with the following parameters [WTS97, Pru95]:

- The constant size of the generated packets is 500 Bytes.
- The average On time for the generator is 72 msec.
- The average Off time for the generator is 105 msec.
- The 'shape' parameter used by the Pareto distribution is 1.35.

Furthermore, we have used an exponential traffic generator which generates the traffic according to an exponential On/Off distribution over an RTP connection for the voice links. Packets are sent (between nodes 0 and 4) at a fixed rate during On periods, and no packets are transmitted during Off periods. Both On and Off periods are exponentially distributed with the following parameters [MBK02]:

- The constant size of the packets is 100 Bytes.
- The average On time is 650 msec.
- The average Off time is 352 msec.
- The average transmission rate during On times in 300 Kbit/sec.

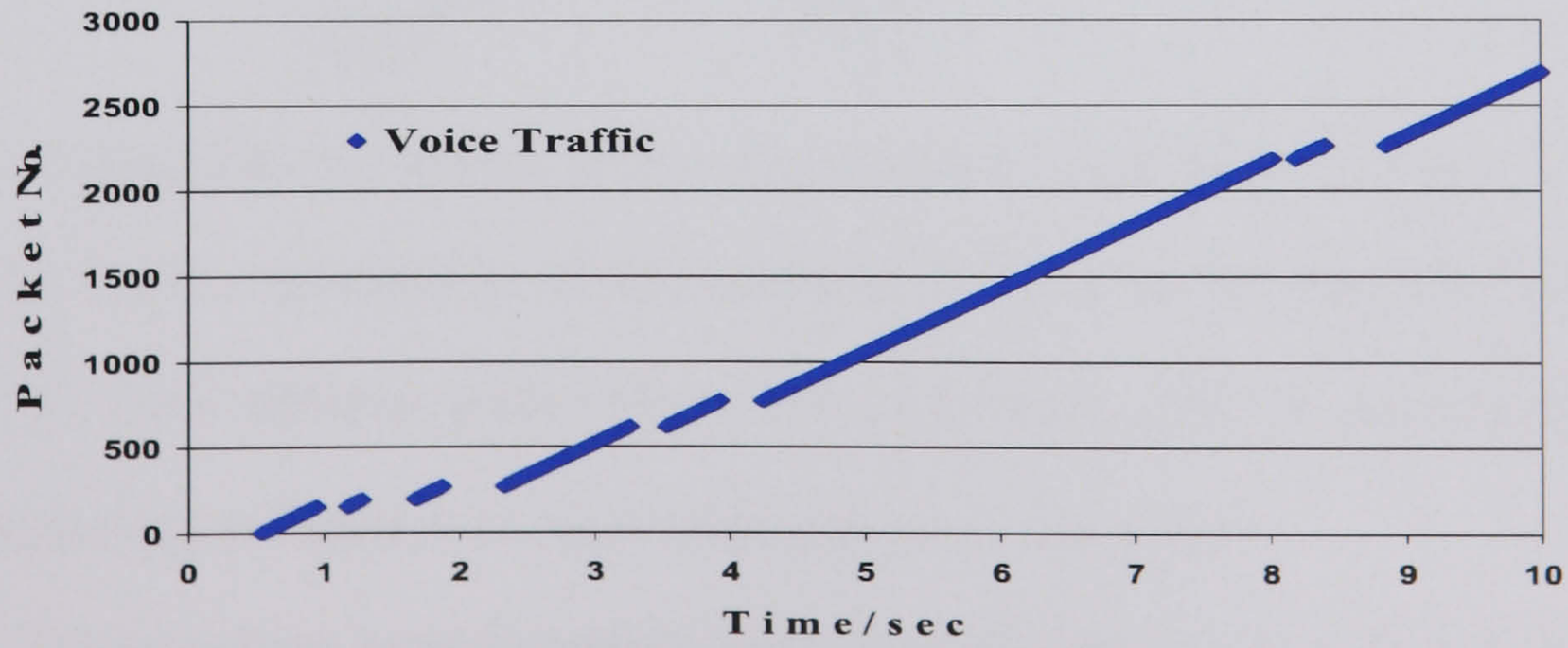


Figure 4-8: Voice traffic with On/Off exponential distribution.

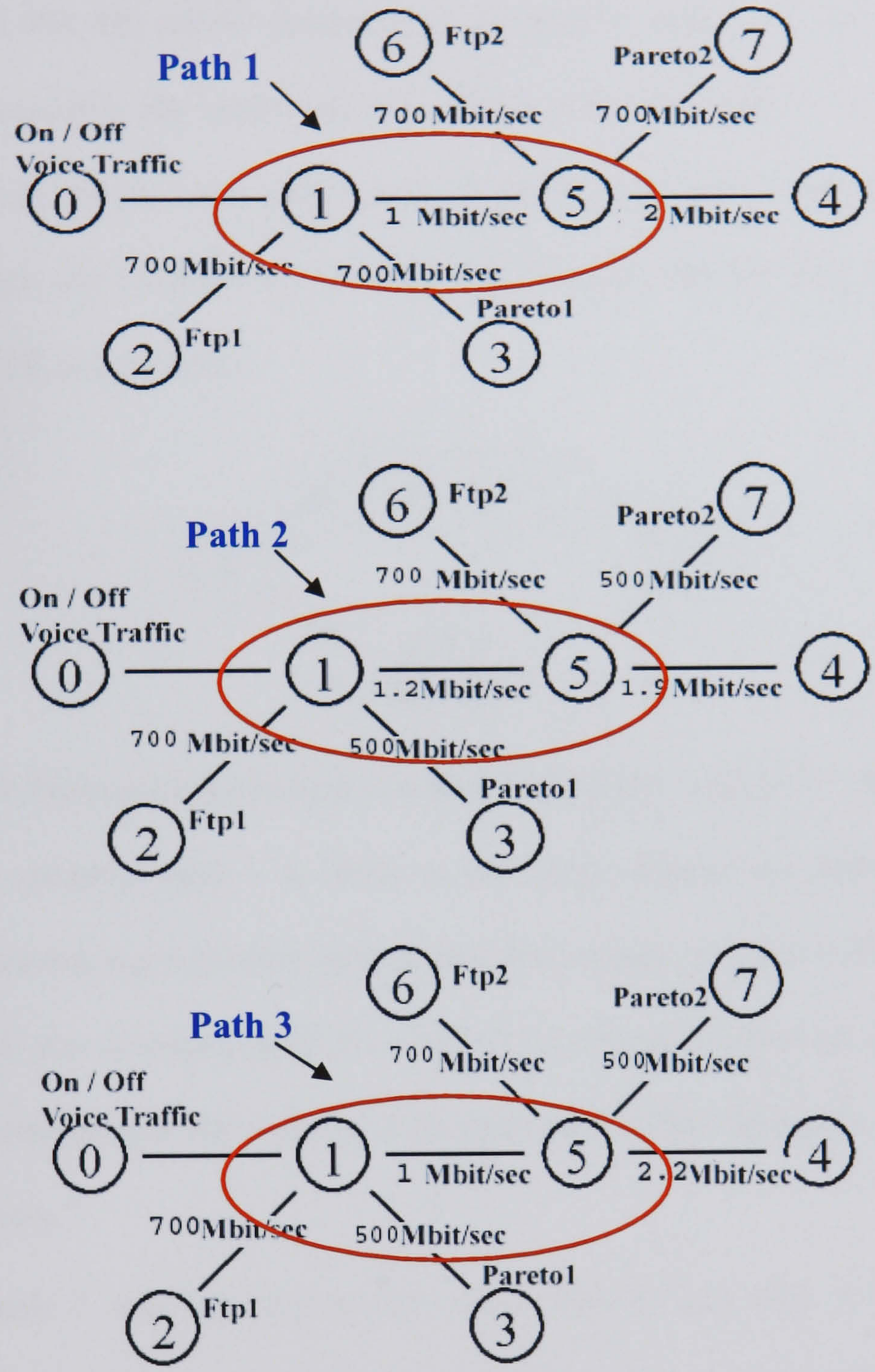


Figure 4-9: The network structure used in simulation for paths 1, 2, and 3.

The voice traffic conveyed by all three paths is chosen to be the same and is shown in Figure 4-8. Three different paths (Figure 4-9), conveying different instantaneous data traffics according to Figure 4-11, were employed in our simulations.

Three scenarios have been simulated based upon the above traffic models and given paths. In the first scenario, each path is considered with no association to any backup

channel and without voice-domain redundancy. A total of 2694 voice packets were transmitted; and 603, 60, and 82 packets were dropped in paths 1, 2, and 3, respectively. In the second scenario, the codec-specific VCBR was used with -1 or -2 redundancy styles (Table 4-2), whilst each path was still considered with no association to other paths. In this case, the numbers of corresponding dropped packets were 369, 24, and 30; or 279, 14, and 18, respectively.

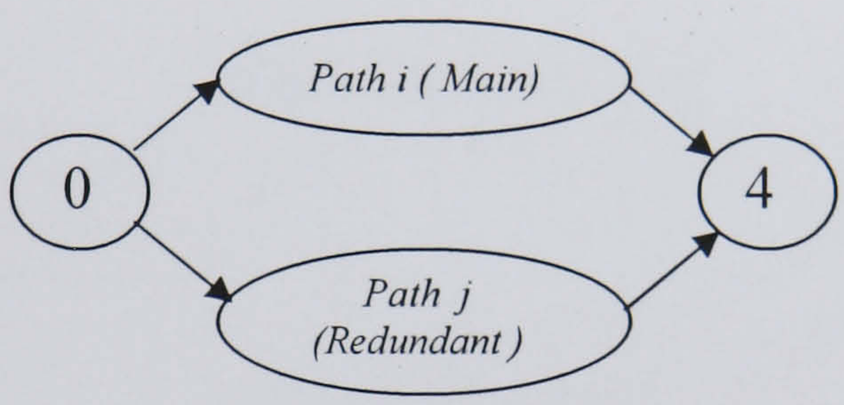
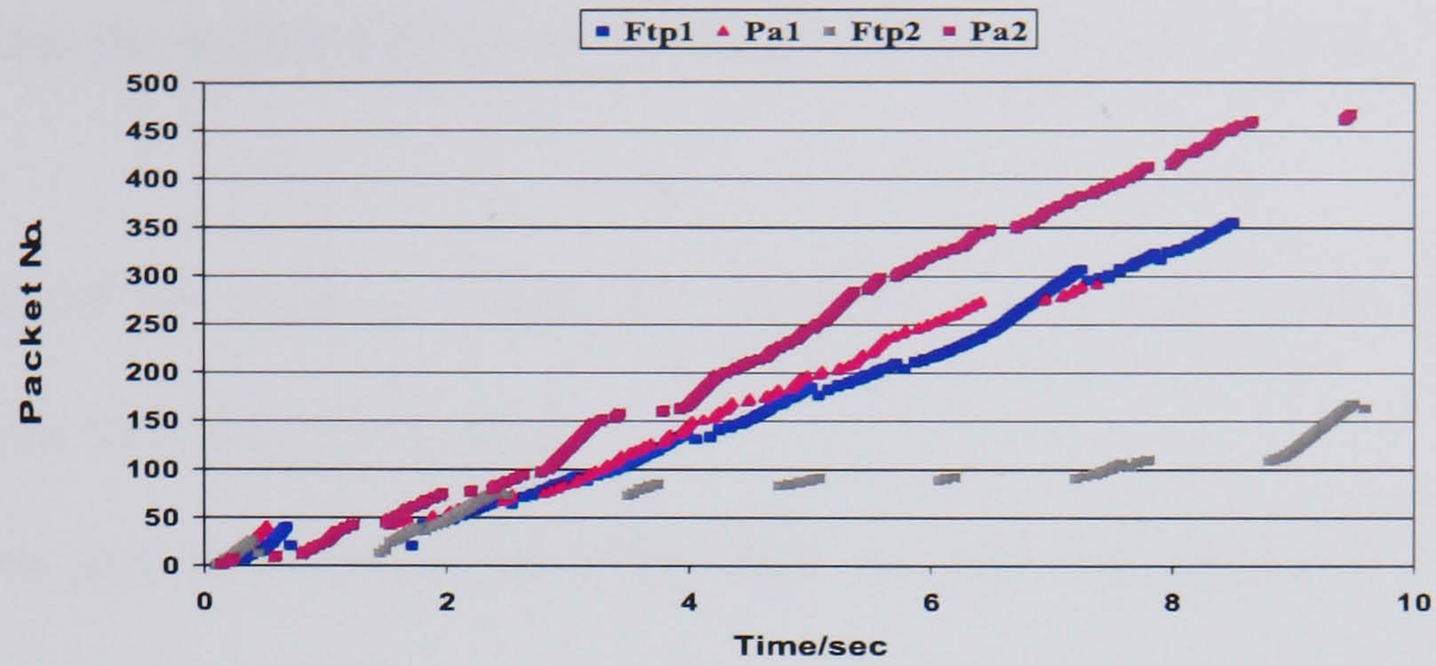


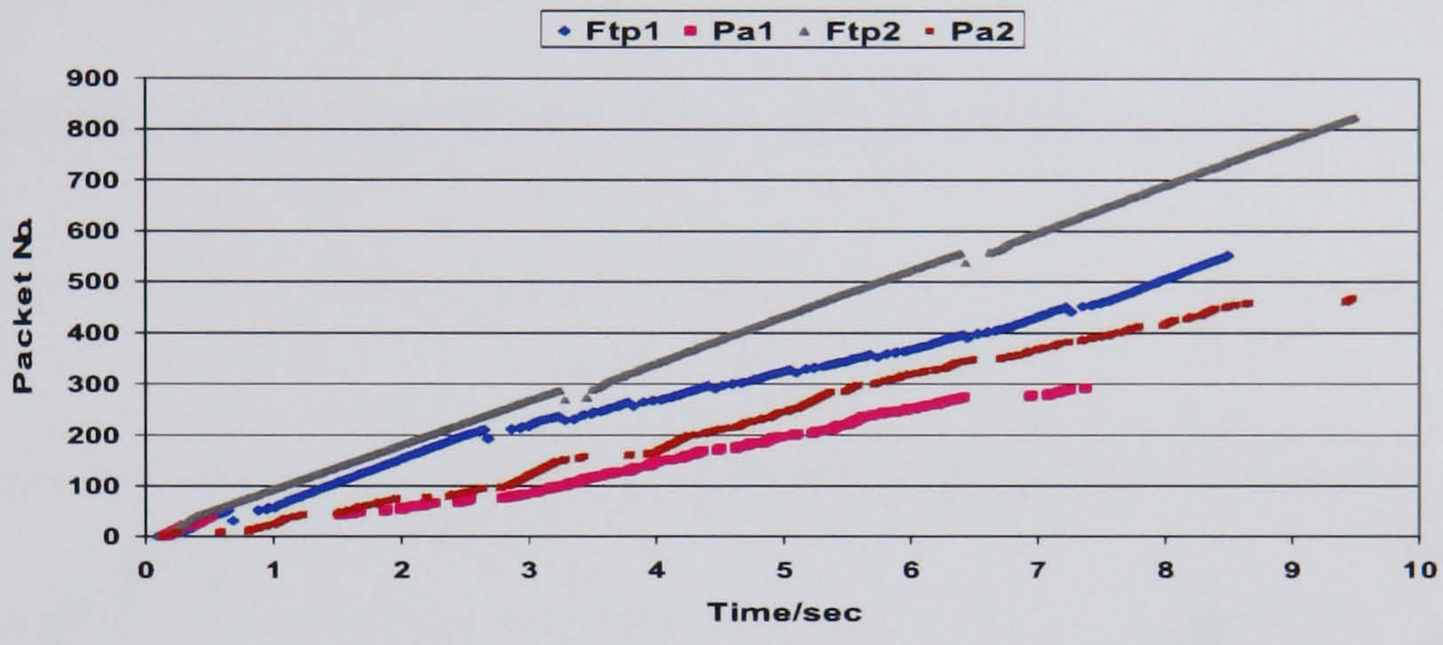
Figure 4-10: Main and redundant path between nodes 0 and 4. (i , j = 1, 2, and 3, $i \neq j$).

In the third scenario, path 1 is taken as the main channel and paths 2 or 3 were employed for conveying redundant information according to Figure 4-10. It can be seen from Table 4-5 that invoking path 2 or path 3 to convey redundant information, has resulted in the reduction in the number of dropped packets through path 1 from 603 to 18 or 22, respectively.

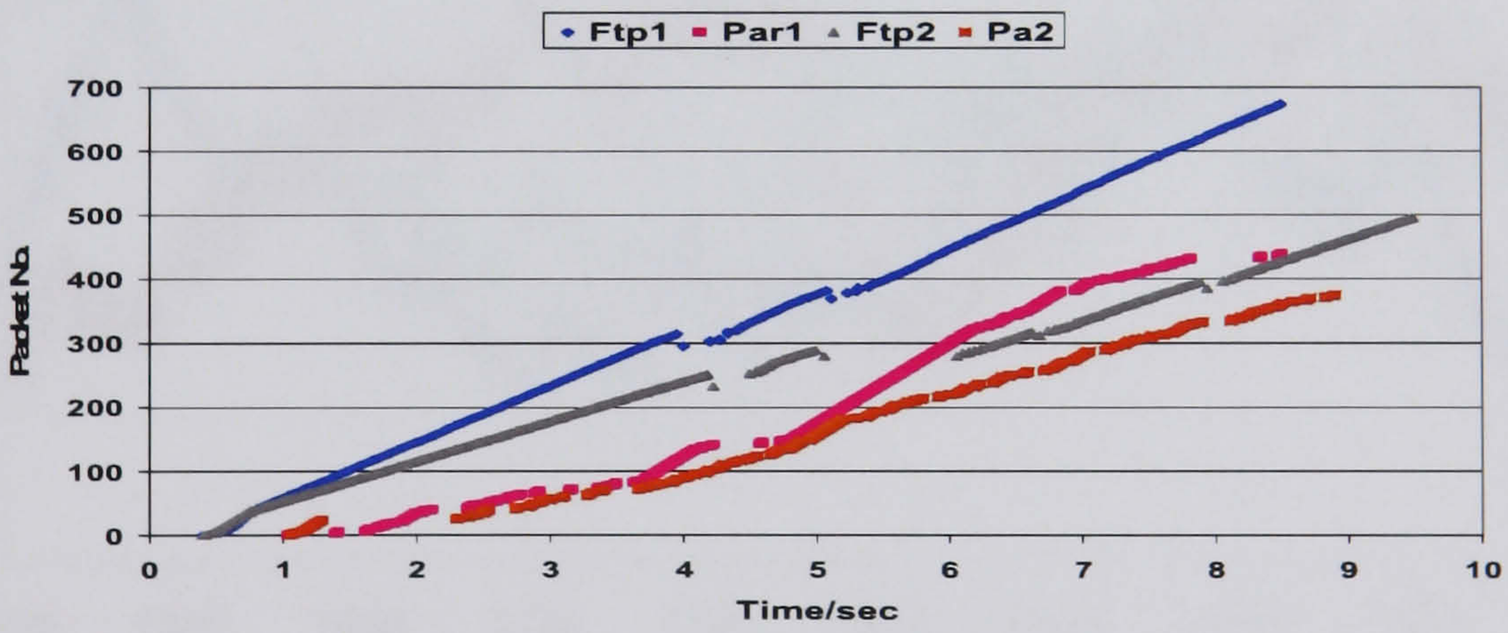
Similarly, path 2 was selected as the main channel and path 3 for sending the redundant information. It can be seen from Table 4-5 that employing path 3 as the backup channel for path 2 has reduced the number of the dropped packets through path 2 from 60 to 4. From Table 4-5, it can be seen that VCBRBC out-performs the codec-specific VCBR, even in the case where one of the paths has an unacceptable condition (path 1).



a) Data traffic for path 1.



b) Data traffic for path 2.



c) Data traffic for path 3.

Figure 4-11: Data traffic for three paths. Ftp1and Ftp2 are two FTP sources of traffic; and Pa1 and Pa2 are two Pareto-distributed sources of traffic.

Furthermore, this technique gives the opportunity to choose between paths for the early packet. As a matter of fact, the VCBRBC not only decreases the loss rate considerably,

but also enables the packet with minimum delay at the receiver to be selected (Figure 4-12).

The selection of the packets which are conveyed by different paths will be made according to the following procedure:

1. If the same packet has been received through all paths, the packet with the least delay will be selected.
2. If a packet has been received only from one path, it should be selected if its respective delay is less than the maximum acceptable play-out delay.

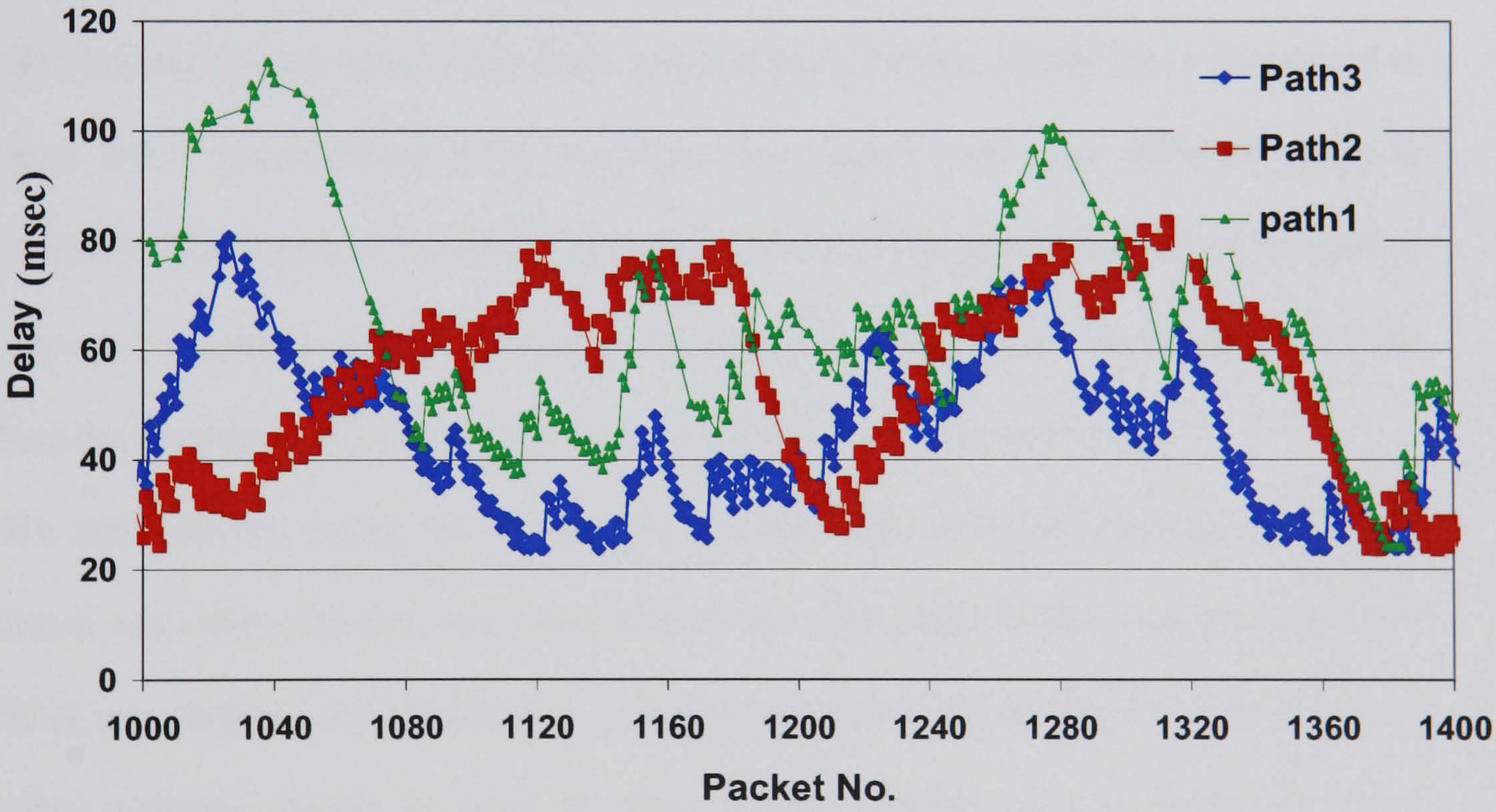


Figure 4-12: Voice packet delay for three paths.

Backup path i / NDP in Main path i	Path 1	Path 2	Path 3	codec-specific VCBR -1	codec-specific VCBR -2
NDP in path 1	603	18	22	369	279
NDP in path 2	18	60	4	24	14
NDP in path 3	22	4	82	30	18

Table 4-5: NDP (No. of Dropped Packets) in three different path.

4.5 Summary

This chapter highlighted our proposed sender-based error correction methods with the analytical and numerical results. Codec-specific VCBR scheme and its improvement in the capacity efficiency and packet loss tolerance, were presented in Section 4.2.1. A new VCBR scheme for real-time voice transmission using a backup channel was introduced in Section 4.2.2. Uncorrelated delay variation and packet loss over different paths is exploited in this method to retrieve the packet received from the backup channels (paths) in case of congestion in the main path. Picking the packet with minimum delay can reduce the average end-to-end delay whilst avoiding late loss or burst loss.

We have shown, using the Gilbert loss model, that over the Internet, these new schemes can reduce the loss rate compared to the single path VCBR schemes, and also without any redundancy, (Section 4.3). Simulation results show that the codec-specific scheme performs almost as good as other VCBR schemes at a considerably lower redundancy overhead, and is also mostly applicable in the IP network in Section 4.4.1; the gain of the backup channel depends on the difference in propagation delay of the multiple paths and the condition of the network traffic in Section 4.4.2

Chapter 5- A New Model for VoIP

5.1 Introduction

Voice over IP between client devices enables users to enjoy the benefits of interactive, conversational communications in a truly multimedia environment. High quality real-time voice communication over the Internet requires low end-to-end delay and low loss rate. The best-effort networks such as the worldwide Internet are, however, characterized by highly varying delay and loss characteristics that cannot comply with today's quality of service requirements. The quality is mainly affected by network impairments such as delay, jitter and packet loss. Playout buffer at the receiving side can be used to compensate for the effects of jitter based on a tradeoff between delay and loss. However, new models for perceived voice quality prediction aim are to find an efficient perceived quality prediction method for perceptual optimization of playout buffer [LI04].

In this chapter, we will introduce a model for a VoIP network using characteristics of queuing delay, delay jitter, dropped packets, and their effect on the network [AS04]. The queuing delay occurs due to the multiplexing of voice packets and data over a shared link. As voice over IP has recently evoked considerable attention, we will give in this Chapter an estimate of the end-to-end delay and dropped loss, experienced for voice packets over IP communication.

Although it is assumed that the model of the network only conveys voice traffic over IP network, and no other Internet data; all effects of the other traffics on the network, over voice traffics, will be considered in the parameters of the model using a self similar

traffic and (appropriate) queuing model. A self-similar and a On/Off exponential traffic model for the IP network and voice traffic, respectively, will be employed to determine the model parameters. The de-jitter buffer length is raised in response to the detection of early congestion through the RED method; whereas a backup channel for transmission of a redundant voice stream is evoked to accommodate a reasonable de-jitter buffer length. We show that the model predicts the overall voice packet loss rate (late and dropped) over the Internet on the BE condition with good precision.

5.2 Traffic and Queue Model

Recent studies have shown that wide-area network traffic is self-similar [Pru95]. Self-similar traffic can be visually characterized by its scale-invariance. Self-similarity implies that the traffic looks the same over any time scale. Self-similarity can be explained as being due to the superposition of many independent and identically distributed (i.i.d), *ON/OFF* sources with infinite variance [ABFRV02]. One well-known heavy-tailed distribution with infinite variance is the Pareto distribution, which has been found to match very well with the actual data traffic measurements.

We note that the Pareto distribution function can be directly derived as a gamma mixture of ordinary exponential densities. With no loss in generality, henceforth we shall use the one-parameter (shape only) version of the Pareto distribution ($\beta = 1$ in equation 3-2) given by

$$p_T(t) = \frac{\alpha}{(t+1)^{\alpha+1}}$$

with n^{th} moment

$$E[T^n] = \frac{n!}{(\alpha - n)!}, \quad n \le a,$$

where α determines the 'heaviness' of the tail in the distribution. For $1 < \alpha < 2$ the mean of the distribution is finite and the variance infinity. The variance becomes finite for $\alpha \ge 2$. When α is closer to 1, the distribution of t becomes 'heavier' and the traffic becomes more bursty. However, it is straightforward to show that the Pareto is indeed a long-tailed distribution. Roughly speaking, α measures the initial rate of decline of the density function curve. A family of such curves is shown in Figure 5-1, together with an exponential curve; which shows the curves with a log scale, to better illustrate their slow rate of decline or heavy-tail nature.

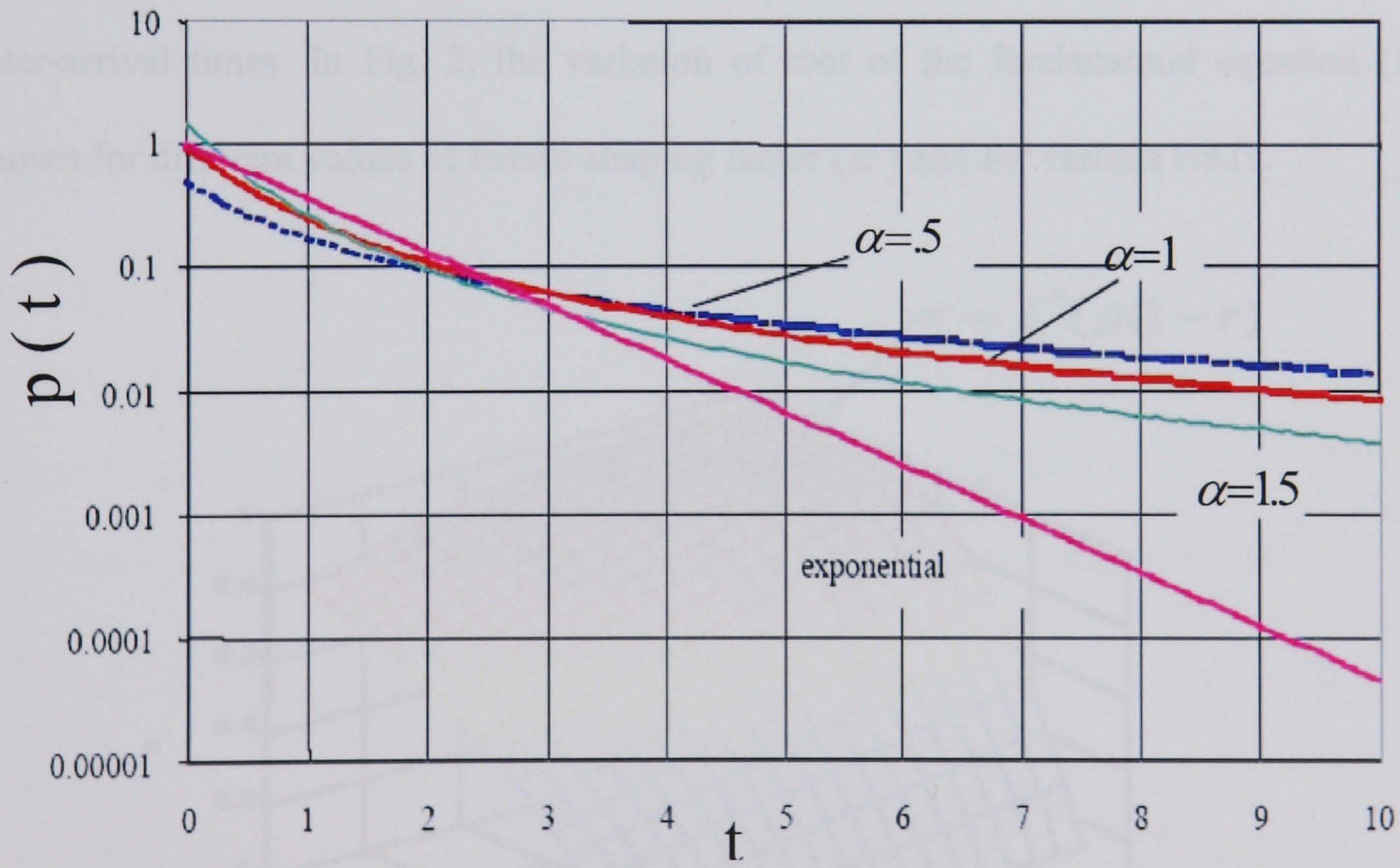


Figure 5- 1: Pareto distribution for several values of α, with exponential distribution.

According to our assumption for the traffic model (a Pareto distribution for the inter arrival times), we are going to employ the standard analysis of a Pareto/M/1/N; queuing system where 'Pareto' indicates the Pareto distribution of the inter arrival times, 'M'

denotes the exponential distribution of the service times with one server, and 'N' is the maximum number of jobs in the system [BGMT98]. The steady-state probability for the number of customers Q_d in the system just before an arrival, is given for all non-negative n by

$$\Pr(Q_d = n) = (1-r)r^n,$$

where r is the root of the fundamental equation (5-1) and $(1/\mu)$ is the expected service time (EST) due to the exponential distribution

$$r = f^*[\mu(1-r)], \tag{5-1}$$

where $f^*(s) = \int_0^\infty e^{-sx} f(x)dx$ is the Laplace transform of the density function $f(x)$ of the inter-arrival times. In Fig. 2, the variation of root of the fundamental equation (1) is shown for different values of Pareto shaping factor (α) and for various ESTs.

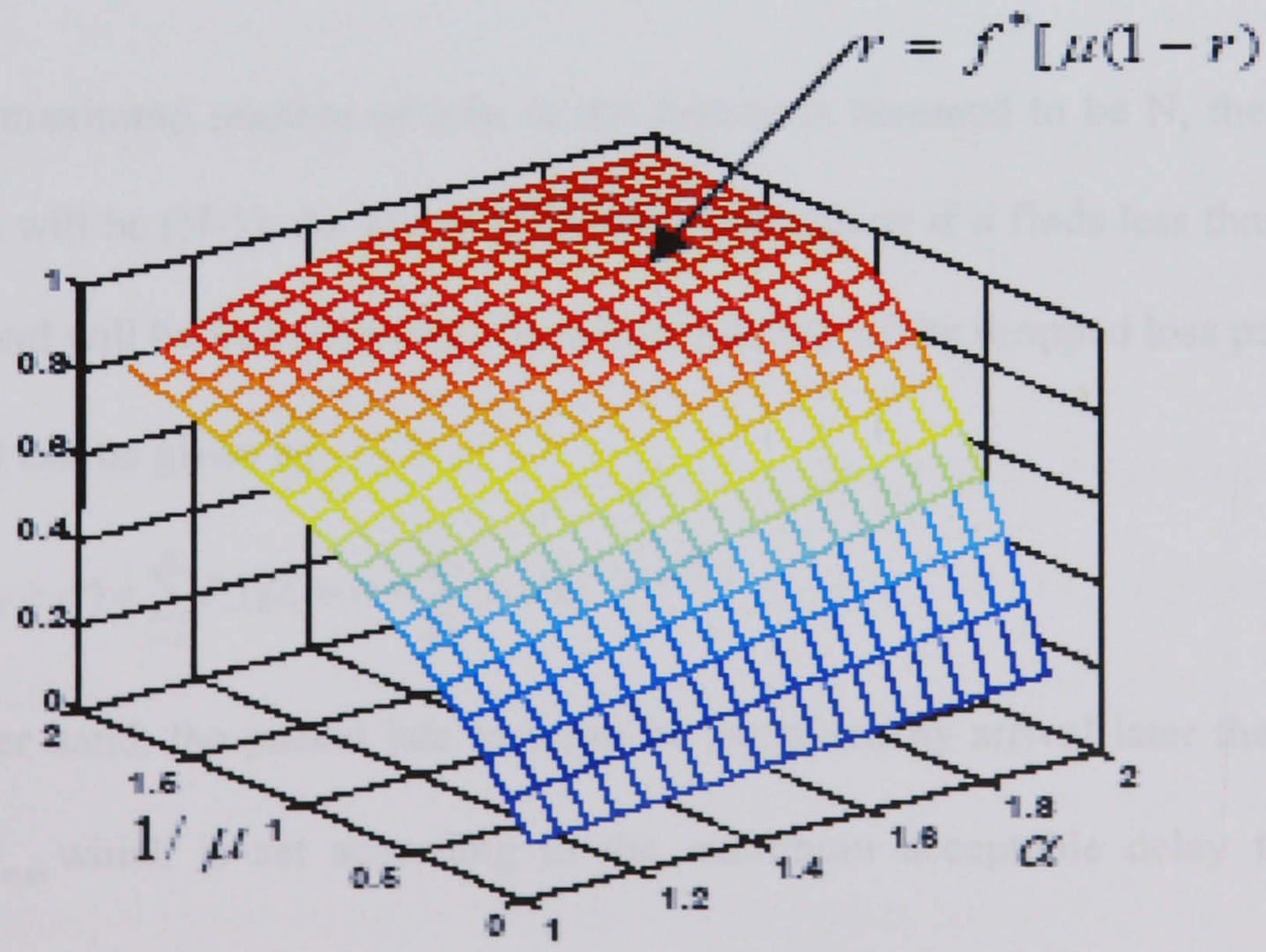


Figure 5-2: r versus Pareto shaping factor (α) and expected service time ($1/\mu$).

In addition, the queue and system waiting-time distribution functions are respectively given as [BGMT98]:

$$W_{qd}(t) = 1 - re^{-\mu(1-r)t} \quad t > 0 \tag{5-2}$$

and

$$W_d(t) = 1 - e^{-\mu(1-r)t} \quad t > 0 \ . \tag{5-3}$$

The mean queue waiting time, D_{qave}, and queue average length, Q_{ave}, according to the Little theory can be written as

$$D_{qave} = \frac{r}{\mu(1-r)} \tag{5-4}$$

and

$$Q_{ave} = \frac{\rho . r}{(1-r)} \ , \tag{5-5}$$

where ρ denotes the utilization factor, namely the fraction of time over which the server is busy.

Since, the maximum number of jobs in the system is assumed to be N, the maximum queue length will be (N-1). An arriving job enters the queue if it finds less than N jobs in the system, and will be lost otherwise. So the probability of the dropped loss packets (P_{DL}) in this model can be given by

$$P_{DL} = P_r(Q_d \geq N) = \sum_{i=N}^{\infty} P_r(Q_d = i) = \sum_{i=N}^{\infty} (1-r)r^i = r^N . \tag{5-6}$$

On the other hand, the packet late loss can be explained by arrival later than a certain threshold $D_{\max}$ which is set according to the maximum acceptable delay for a voice

communication. According to the Pareto/M/1 queuing model, the late loss probability P_{LL} can be defined as

$$P_{LL} = P(t > D_{Q\max}) = \int_{D_{Q\max}}^{\infty} d(W_{qd}(t))dt$$

$$= \int_{D_{Q\max}}^{\infty} (r\mu(1-r)e^{-\mu(1-r)t})dt = re^{-\mu(1-r)D_{Q\max}}, \quad (5\ 7)$$

where $W_{qd}(t)$ is the PDF of the queue waiting time, given by equation 5-2; and $D_{Q\max}$ represents the maximum queuing delay percentile incurred in the network, given by 5-9.

Delay is measured end-to-end across the packet network from the point where the voice is coded at the source, to the point where it is decoded at the destination [Tog99]. End-to-end delay (EED) will be comprised of two elements: fixed and variable delay. Fixed delay components are encoder delay (D_{enc}), packetization delay (D_{pack}), serialization delay (D_{ser}), propagation delay (D_{pro}) and decoder delay (D_{dec}). Variable delay arises from queuing delay D_Q which is experienced by data waiting in the buffers to be served by the resources within the network. These buffers create variable delays across the network. This effect is known as delay jitter and is handled by the de-jitter buffer at the destination.

In order to absorb the variability in the delay between one packet and another, a de-jitter buffer is implemented at the destination. When packets arrive, they are not played out immediately, but are stored in a buffer. The voice will be played out in the destination only when a sufficient number of frames is available in the buffer. In other words, the de-jitter buffer transforms the variable delay into a fixed delay by holding the first received

sample for a period of time before playing it out. This holding period is known as the initial play out delay (POD_{min}).

Proper handling of the de-jitter buffer is a critical task. If samples are held for too short a time, delay variations may cause the buffer to under-run and to produce silence gaps in the speech. If samples are held for too long, the buffer can overrun, and the dropped packets again cause silence gaps in the speech. Lastly, if packets are held for too long, the overall delay on the connection may rise to unacceptable levels. Delay budget can be given by the following equation

$$(D_{enc} + D_{pack} + D_{dec}) + \sum_{Hops}(D_{ser} + D_{pro} + D_Q) + POD \le D_{max} \quad , \tag{5-8}$$

where D_{max} is the maximum acceptable delay for a voice communication, and the summation is over the numbers of hops in the network. Despite the second term in (5-8), the first term is not dependent on the path nor on the number of hops. So the second term is expected to be subjected to more variations with the network conditions. For a given voice connection, the only random component in voice delay, that is the only source of jitter, is queuing delay in the network. The play out delay, POD, insures that most of the transmitted packets are available to the decoder at an appropriate time. Assuming that D_{Qmax} represents the maximum queuing delay percentile incurred in the network, the receiver must delay the first packet of a voice stream by D_{Qmax}, i.e.

$$POD = \sum_{Hops} D_Q = D_{Q\max} \quad .$$

Assuming that the packet has already been subjected to a queuing delay equal to D_{Qmax}, the end-to-end delay budget equation becomes

$$D_{Q\max} = \sum_{Hops} D_{q\max} \le [\mathrm{D}_{\max} - (\mathrm{D}_{\mathrm{enc}} + \mathrm{D}_{\mathrm{pack}} + \mathrm{D}_{\mathrm{dec}}) - \sum_{Hops} (\mathrm{D}_{\mathrm{ser}} + \mathrm{D}_{\mathrm{pro}})] \cdot \quad (5\text{-}9)$$

5.3 New VoIP Network Loss Model

The Internet is designed for BE datagram services with no assurance for actual packet delivery. Since there is no dedicated end-to-end connection between the sender and the receiver, occurrence of packet loss, out-of-order delivery, delay jitter, and latency are inevitable in cases where the shared network is congested. Detailed dynamic analysis of an IP network such as the Internet requires deterministic or statistical knowledge of the instantaneous traffic conveyed by the system. Acquiring this information, be deterministic or statistical, can be an extremely complicated task especially in networks that support various classes of traffics and QoS.

For many of the design and analysis purposes, however, it suffices to obtain a reliable estimate of only a number of key performance indicators. Probability of voice packet loss can be viewed as one of such performance indicators in VoIP networks. This factor is a combination of probabilities of late loss P_{LL} and dropped loss P_{DL}.

These two constituent elements of loss were derived in Section 5.2, and we invoke them in this section to propose a simplified model for analyzing VoIP networks. In this model, shown in Figure 5-2, the status of the network at a given instant is represented by the values of P_{DL} and P_{LL} at that moment.

Experimental results have shown that the quality of voice decreases considerably should the voice packet loss exceeds 5% [Rin99]. As we mentioned earlier, a voice packet is considered as a lost packet in both cases of being dropped or delayed for more

than $D_{Q\max}$ (seconds). To analyze the late loss and dropped loss for voice packets in our model, the distribution function of the voice delay and dropped packet loss should be known or estimated. The late and dropped loss probability can be calculated by using the value of *r*, the root in equation (5-2); and it is in turn, obtained from the knowledge of the average voice packet delay time, equation (5-5), over a given time interval. There are efficient methods such as random early detection (RED) and RTCP to derive reliable estimates of the average voice packet delay, and thus the value of *r*.

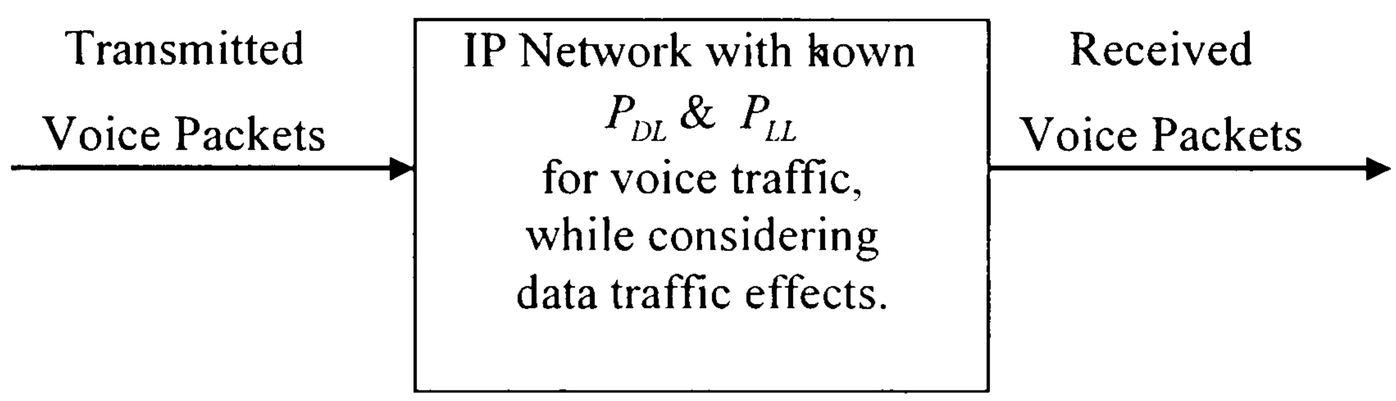


Figure 5-3: IP network as a black box with given voice dropped loss P_{DL}, and late loss P_{LL}, probability.

Random early detection (RED) as an effective mechanism to control the congestion in the network routers and gateways can be used to measure the traffic load level in the queue by invoking the average queue size (discussed in Section 2.8.3). This is calculated using an exponentially weighted moving average filter and can be expressed as

$$avg_n = (1-w_q)\, avg_{n-1} + w_q . q$$

where q is the actual queue size, w_q is the filter weight which determines the time constant of the low-pass filter, and avg_n denotes the estimate of the average queue size at step n. The probability of dropping a packet arriving at the queue depends on the average

queue length, the time elapsed since the last packet was dropped, and the maximum dropping probability parameter (max_p). If the average queue size is larger than a maximum threshold (max_{th}) all arriving packets are dropped.

As discussed in Section2.2, RTP has the responsibility for recovering lost segments and resequencing of the packets for the application layer. The accompanying RTCP provides feedback of the quality of the data delivery and information about session participants. The format of the RTCP packets is fairly similar to RTP packets. The main function of the RTCP is, QoS monitoring, congestion control, identification, and session size estimation and scaling.

To verify the usefulness of this model, we need to show that the mentioned parameters are actually the key performance indicators for a VoIP network. That is, we need to show that the knowledge of P_{DL} and P_{LL} provides the essential information about the QoS offered by the network. Compliance of the proposed model with this requirement is obtained by the fact that quality of voice at the receiver is mainly affected by the number of over-delayed packets and that of lost packets.

5.4 Numerical Result

In this section the simulation and numerical results are obtained to evaluate the network model and demonstrate the application of the model for network analysis. In the simulations we have considered a linear topology, depicted in Figure 5-3, for the IP network. There are two congested links between the edge and the core routers, namely nodes zero and seven, which are shared by data flows. The capacitys of the two congested links are L_1=1 Mbps and L_2=1.4 Mbps for the path condition 1, whereas L_1=

0.8 Mbps and L_2= 2 Mbps for the path condition 2. The data traffic is produced by two FTP sources, Ftp1 and Ftp2, and two Pareto-distributed traffic generators, Pareto1 and Pareto2. The traffic is conveyed by the congested links within the path. Also the voice traffic aggregates consist of packets generated by sources which use the same codec scheme.

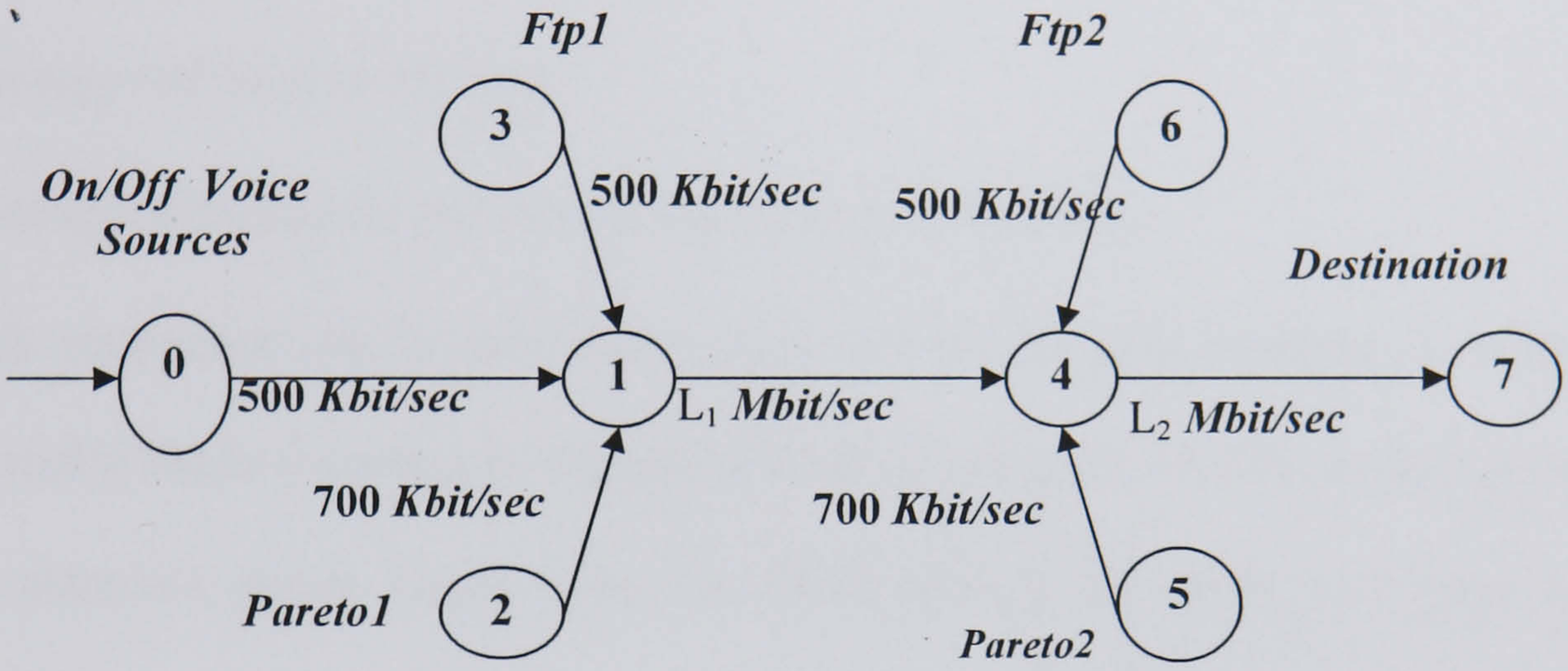


Figure 5-4: The network structure used in simulations.

We have chosen an FTP source over the TCP connection between nodes 3 and 7 and between nodes 6 and 7 (Figure 5-3). The On/Off Pareto distribution is used between nodes 2 and 7, and between nodes 5 and 7, with the following parameters [WTS97, PRU95]:

- The constant size of the generated packets is 500 Bytes.
- The average On time for the generator is 72 msec.
- The average Off time for the generator is 105 msec.
- The 'shape' parameter used by the Pareto distribution is 1.35.

Furthermore, we have used an exponential traffic generator which generates the traffic according to an exponential On/Off distribution over a RTP connection for the voice

links. Packets are sent between nodes 0 and 7 at a fixed rate during On periods, and no packets are transmitted during Off periods. Both On and Off periods are exponentially distributed with the following parameters:

- The constant size of the packets is 100 Bytes.
- The average On time is 650 msec.
- The average Off time is 352 msec.
- The average transmission rate during On times is 300Kbit/sec.

Network congestion can be seen from Figure 5-6 for the path condition 1; with voice and data traffic loads according to Figures 5-4 and 5-5 respectively.The actual queue size and the estimated queue length from the RED method are given in Figure 5-6 for comparison.

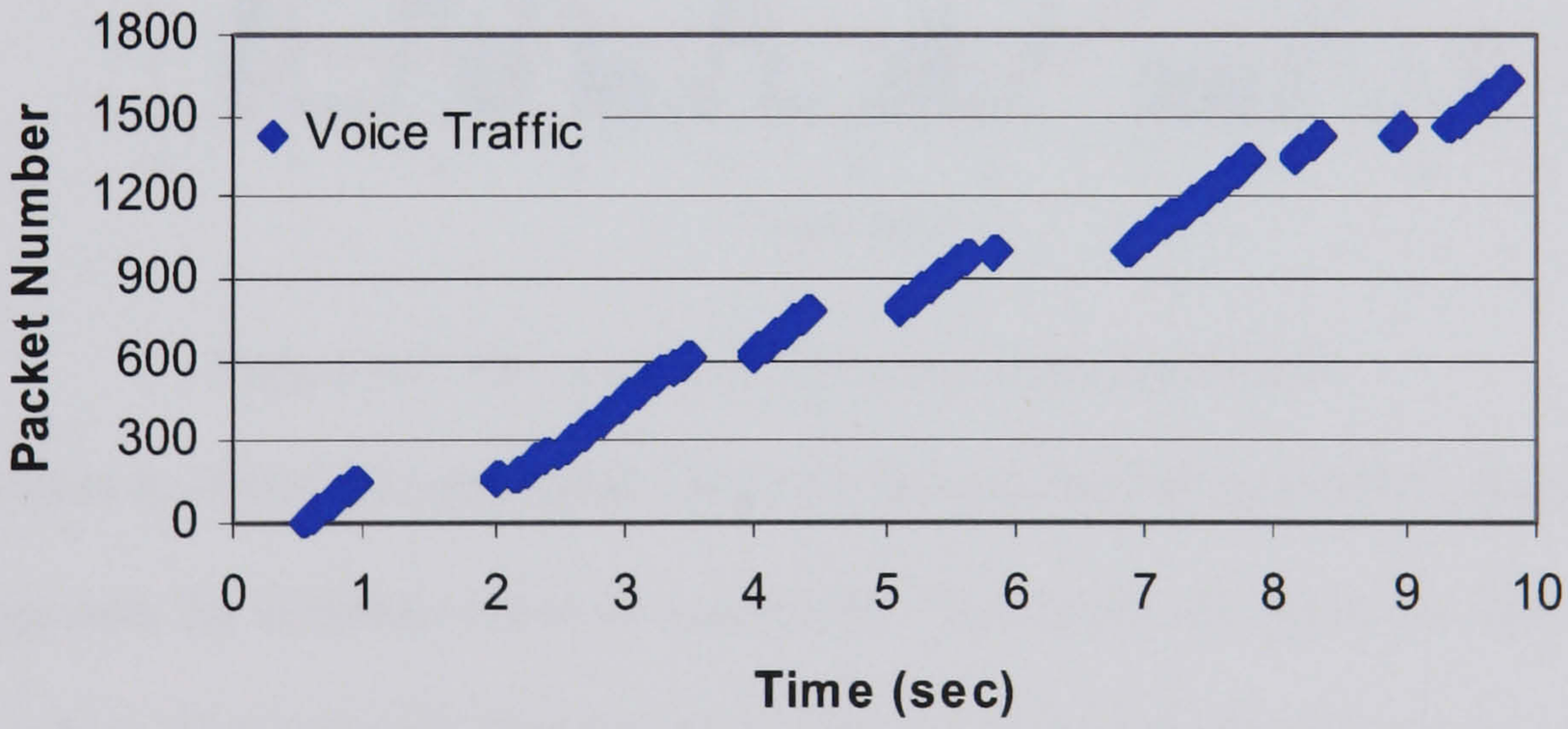


Figure 5-5: Voice traffic with On/Off exponential distribution.

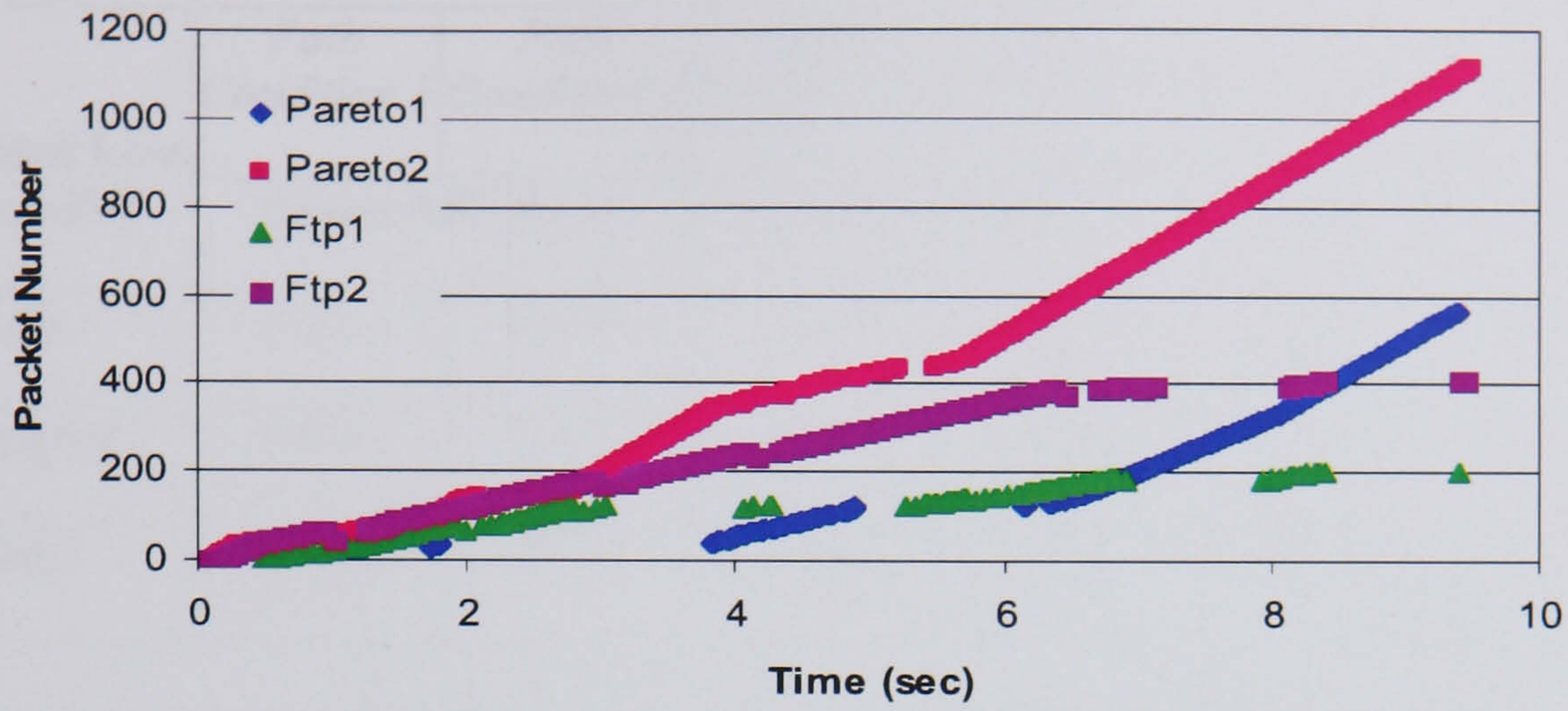


Figure 5-6: Data traffic with two ftp sources Ftp1 and Ftp2, and two Pareto-distributed sources Pareto1 and Pareto2.

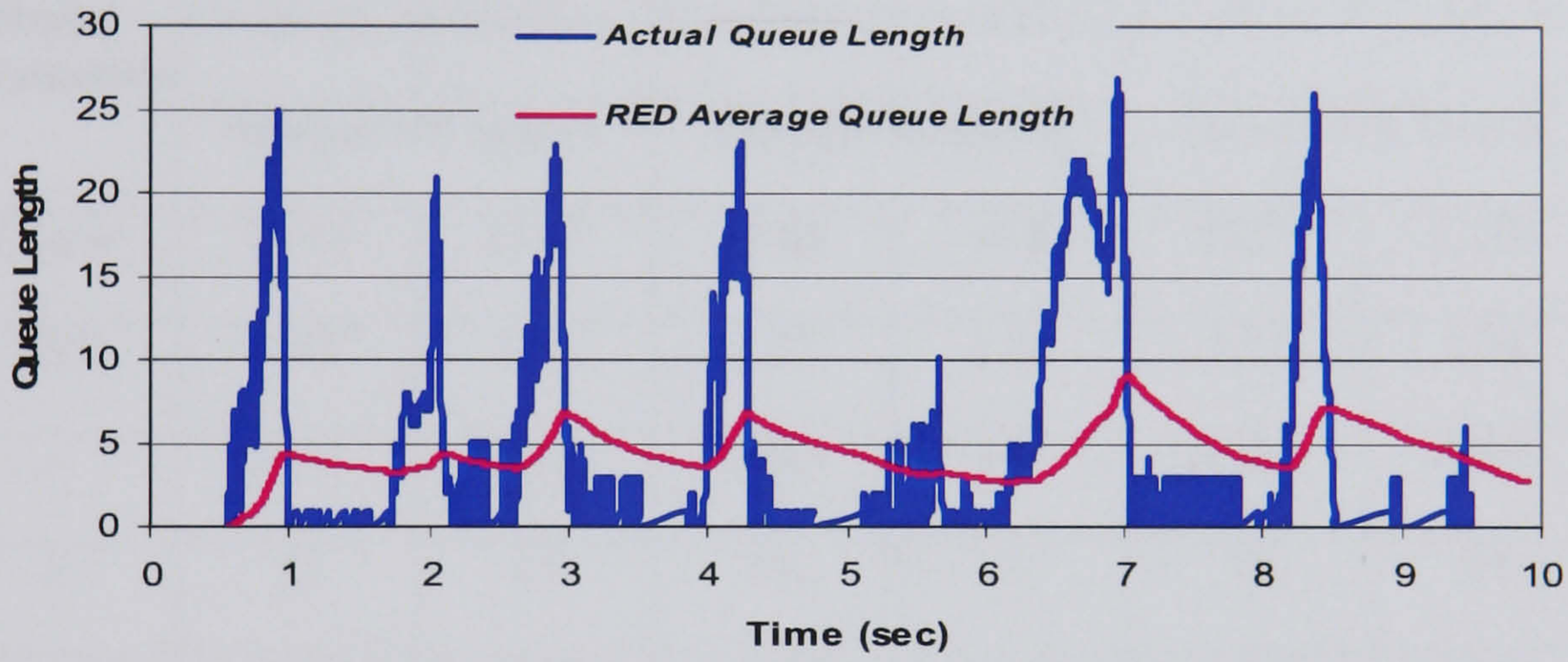


Figure 5-7: The average (RED) and actual queue length.

NS2 is used to obtain the numerical values for the number of transmitted, dropped, and late loss packets for different network conditions. The results are given in Table 5-1 for different values of maximum play out delay D_{Qmax}. By increasing the length of the de-jitter buffer from 40 msec to 60 msec, the network late loss can be improved; and the loss probability of the network decreases from 0.176 to 0.142 for path condition 1, and from 0.140 to 0.076 for path condition 2 (Table 5-1).

Simulated Loss Probability	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*
	***Dqmax*=40 (msec)**		***Dqmax*=50 (msec)**		***Dqmax*=60 (msec)**	
Late	0.101	0.089	0.090	0.035	0.067	0.025
Dropped	0.075	0.051	0.075	0.051	0.075	0.051
Total	0.176	0.140	0.165	0.085	0.142	0.076

Table 5-1: Simulation result for path condition 1 and 2 for different values of D_{qmax} .

Loss Model Parameters	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*
	Dqmax=40 (msec)		**Dqmax=50 (msec)**		**Dqmax=60 (msec)**	
Dqave	16.41	15.57	13.05	14.8	12.75	12.40
EST	1.37	1.34	1.37	1.34	1.37	1.34
r	0.923	0.92	0.905	0.91	0.903	0.902
n	32	35	26	31	25	29
a	1.5019	1.509	1.453	1.483	1.448	1.462

Table 5-2: Loss model parameters for path conditions 1 and 2 for different values of D_{Qmax}, given the mean queue waiting time (*Dqave*) and the expected service time (EST).

The model parameters obtained, based upon the analytical results of Sections 5.2 and 5.3, are given in Table 5-2; and the theoretical late and dropped loss probabilities are shown in Table5- 3.

Analytical Loss Probability	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*	*Path Condition 1*	*Path Condition 2*
	Dqmax=40 (msec)		**Dqmax=50 (msec)**		**Dqmax=60 (msec)**	
Late	0.0975	0.0620	0.0556	0.0317	0.0316	0.0162
Dropped	0.090	0.059	0.090	0.059	0.090	0.059
Total	0.187	0.121	0.145	0.091	0.122	0.075

Table 5-3: Analytical results for path condition 1 and 2 for different values of D_{Qmax}.

In Figure 5-7 the simulation results for loss probability are given together with the analytical results obtained from our proposed method. It can be seen that the model can predict the network behavior with good accuracy.

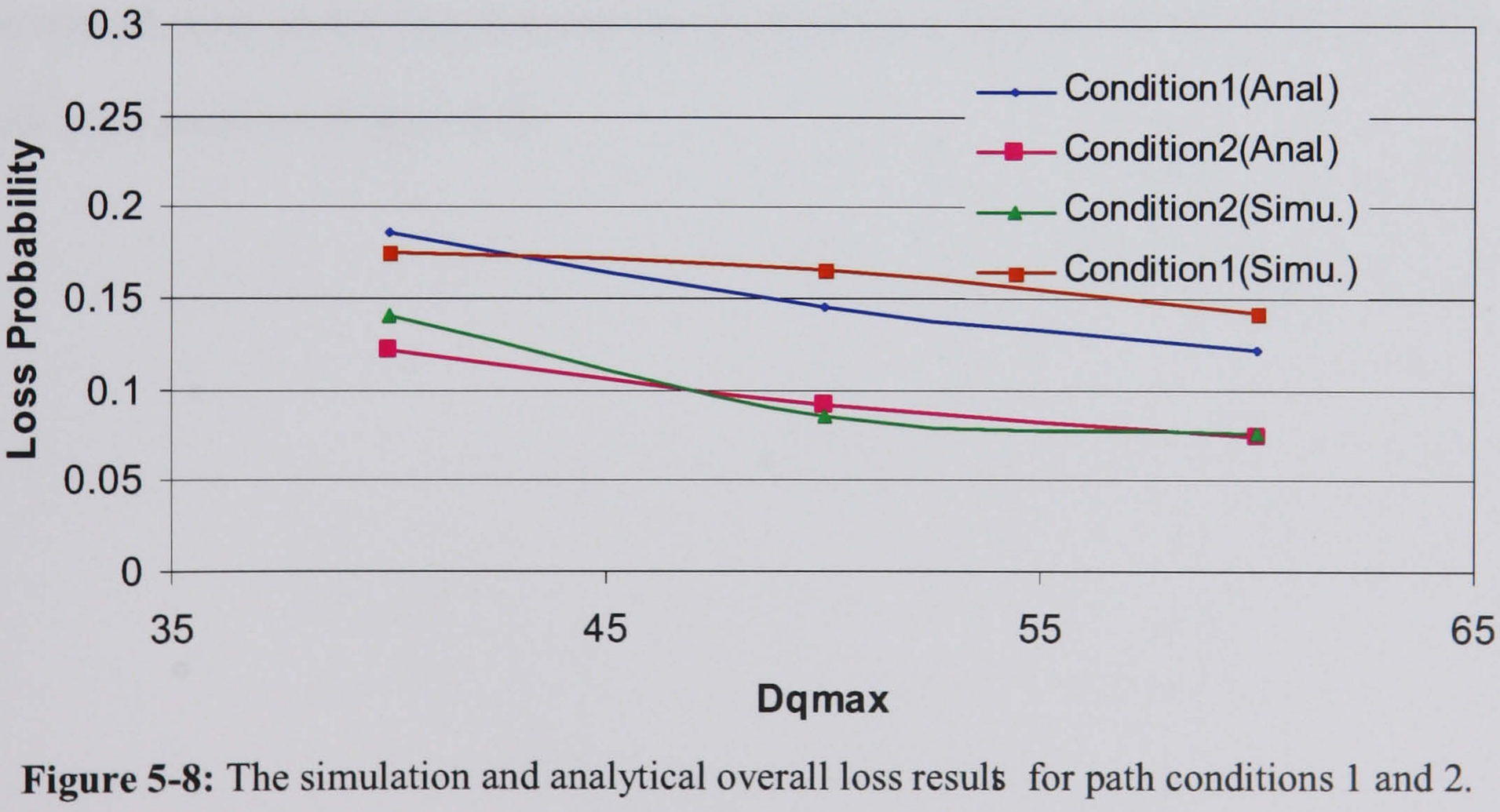


Figure 5-8: The simulation and analytical overall loss results for path conditions 1 and 2.

The accuracy of the model can be further enhanced by using improved methods for traffic estimation and a better queue model.

5.5 Summary

In this chapter we introduced a loss model using characteristics of queuing delay, delay jitter, and number of dropped packets; and analyzed their effect on the performance of VoIP networks. The traffic and queue model were explained in Section 5.2. In this section a self-similar and On/Off exponential traffic model for the IP network and for voice traffic respectively, have been employed to determine the model parameters. In Section 5.3 we introduced the new model for VoIP networks. In this model the status of the network at a given instant has been represented by the values of probabilities of late loss P_{LL} and dropped loss P_{DL} at that moment. The proposed model is numerically evaluated in Section 5.4. The accuracy of the model is verified through simulation and analytical results for different traffic conditions; and it is shown that the model predicts the overall voice packet loss rate (late and dropped) over the Internet on the BE condition with good precision (Figure 5-7).

Chapter 6- Voice over Adaptive IP Networks

6.1 Introduction

Utilizing packet data networks and the Internet in particular, to transport voice and fax traffic that traditionally run on the circuit-switched and PSTN, has attracted much interest recently. Voice over IP between client devices enables users to enjoy the benefits of interactive, conversational communications in a truly multimedia environment. High quality real-time voice communication over the Internet requires low end-to-end delay and low loss rate. The BE networks such as the worldwide Internet are, however, characterized by highly varying delay and loss characteristics that cannot comply with today's QoS requirements. The overall objective in transmission over IP networks is to find an array of technical solutions to guarantee the desired level of the QoS under most network conditions [Rin99].

Packet loss in delay-sensitive applications such as interactive VoIP is a result of not only packet erasure, but also delay jitter [ASC04]. Due to the stringent delay budget and the need to output speech periodically and continuously, packets experiencing sudden high delays have to be discarded at the receiving end if they arrive later than the scheduled play out deadline, which results in late loss.

Differentiated Services (DS) approach is proposed as a scalable QoS solution for the Internet [WM01]. To provide a delay bound for the real-time packets, it is isolated from the TCP traffic. The real-time services require limited buffer size in the routers so that the

packet could have a total bounded delay. We propose a selective dropping mechanism to make the deadline misses distributed evenly during congestion period, so the voice quality is degraded gracefully during congestion.

The adaptive error correction method in VoIP application has several appealing features. Firstly, an efficient use of the network resources is made, since adaptive applications do not need a rigid partitioning of the link bandwidth. Secondly, it can exploit better QoS than the plain applications. Thirdly, by adapting gracefully to network conditions, it leaves more resources for signaling and critical in-band flow management sharing the same network facilities [BCDM01]. Also, adaptive application in the Internet would reliably achieve the ability of making a trade-off between throughput, QoS and utilization of the network.

However, another approach to improve the QoS is taken by employing AVoIP systems, where the rate of the single speech sources is dynamically adapted to the workload conditions. A variable bit rate (VBR) speech coder chooses the most appropriate bit rate from a predefined set of operating modes: source or network-driven [EH99]. To guarantee a certain QoS, even in critical conditions featuring great delays and background noise levels, it is necessary to control the peak rate, and therefore use a multi-rate codec. It is also necessary to have proper comfort noise models that only multimode coding can provide.

In this chapter, we will focus our attention on providing adaptability to IP telephony applications, through variable bit-rate coding algorithms; and use a selective dropping mechanism to make the deadline misses distributed evenly during congestion. Moreover, an adaptive VCBRBC scheme is proposed on the basis of estimation of the network

conditions, measured in terms of de-jitter buffer length; and using the average queue size in RED to control the amount of redundancy, or the need for making use of a backup channel, more efficiently. These algorithms aim to control the load of the network, and use the network resources efficiently.

6.2 Adaptive Rate/Error Correction Algorithm

As explained in chapter 2, G.729 at 8kbit/s is a recent good-quality ITU-T speech coding standard based on CS-ACELP. More recently, ITU-T has standardized two extensions of the 6.4 Kbit per second and 11.8 Kbit per second, respectively indicated as G.729 annex D and E. These two extensions have been a significant reference point for the development of the hybrid multimode multi-rate codec. The bit allocation for G.729 and its extension are summarized in Table 6-1 [ITU00].

Parameters	G.729	Annex D	Annex E forward	Annex E Backward
LPC	18	18	18	
Pitch period	13	12	13	13
Parity bit	1	0	1+1+1	1+1+1
Codebook	34	22	70	88
Pitch Code book gain	14	12	14	14
Total Bits	80	64	118	118

Table 6-1: Bit allocation, every 10 ms, for G.729.

The proposed error correction control algorithm is summarized in Figure 6-1. The network loss condition is estimated, at the beginning, in accordance with RTCP reports; and tries to relieve the network congestion by changing the codec rate for several voice sources. The basic idea of the proposed adaptive algorithm is that according to the packets delay and loss on the network, estimated by the RTCP receiver reports at the beginning and by the RED scheme after enabling the DS; the voice coder rate should be changed.

The adaptive approach has arisen on the following concept: at the beginning of the congestion period the network is not mainly affected, so according to the algorithm, the RTCP report would be enough to present the network condition. For the normal network condition, low loss and delay, just the algorithm checks if it is possible to use the increased codec bit rate for the better quality. However, if the packet loss or delay, missed their tolerable region by more than 5% packet loss or more than 150 millisecond delay (for the end-to-end one way delay according to the recommendation G.114 by the ITU), the DS scheme would be enabled to prepare a more powerful routing and monitoring on the network. After enabling the DS, the average queue size in the RED algorithm will be used to assess the network condition and decide if it is time to increase or decrease the codec bit rate.

Furthermore, an adaptive approach provides the ability for choosing the optimal voice codec bit rate of the current instantaneous channel conditions. After enabling the DS, the network congestion and traffic load level are detected using the average queue size (avg) in the RED algorithm. However, the adaptive approach makes it possible to run the best

trade-off between voice codec bit-rate, packet loss and delay; and thus to operate, at least in principle, at the instantaneous optimal operating point.

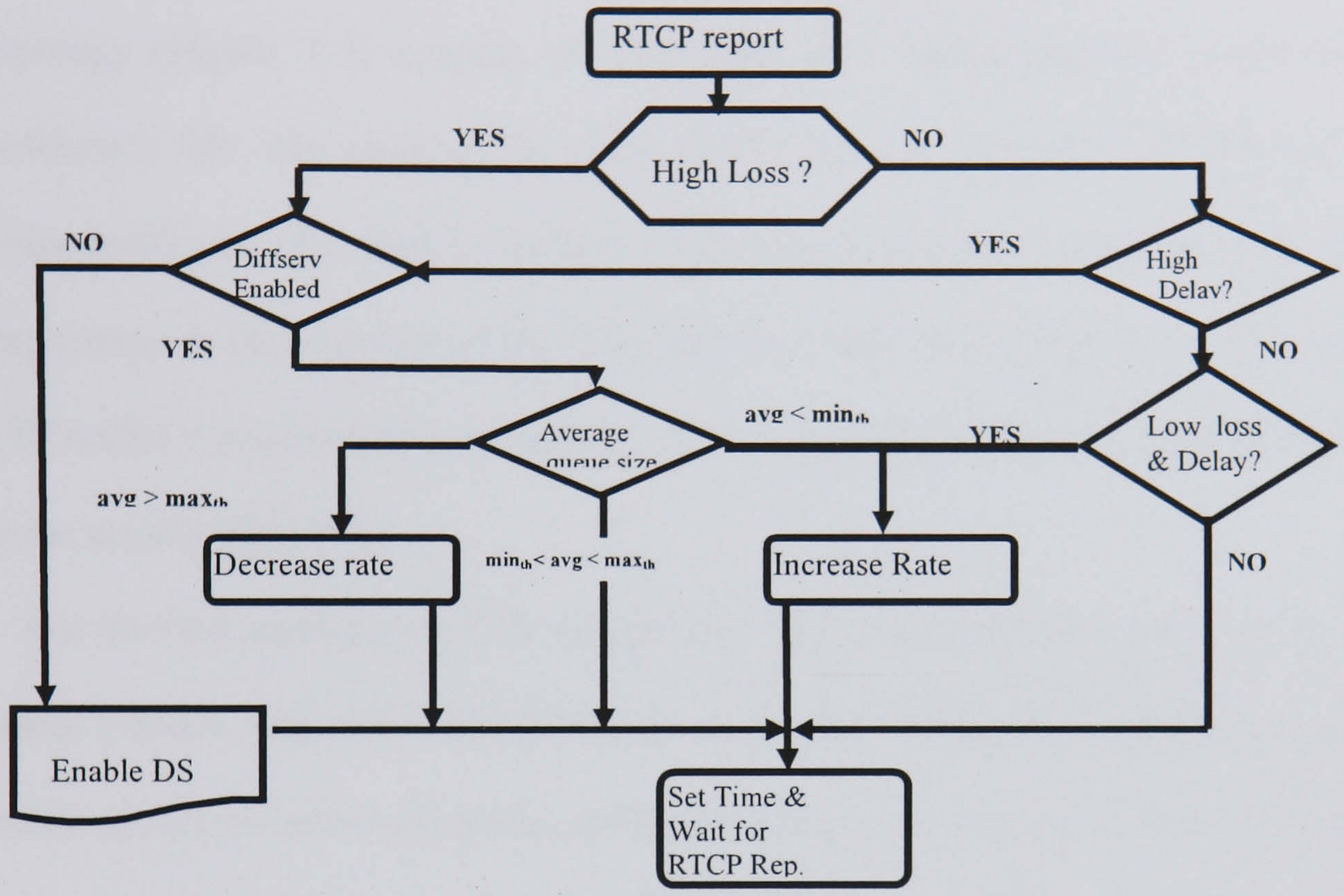


Figure 6-1: Adaptive rate/error control algorithm.

In order to demonstrate the effectiveness of our algorithm, we use simulation results to compare the performance of simple drop-tail (FIFO) queuing, with that of RED. Since the focus in our simulations is to show how efficiently the queue policy supports the UDP traffic as well as the TCP traffic, we only consider the QoS for voice class and BE class. Note that our approach can be extended to support more classes by using multiple buffer occupancy thresholds.

6.2.1 Simulation Result

In this section (using the network simulator NS2) we have considered a single congested link between edge and core routers, which is shared by all flows. The linear topology (Figure 4-5) consists of core links of 1 Mbps capacity, which act as the bottleneck link. The topology is constructed to provide four main 300 Kb/Sec links for voice traffic (as discussed in Section 4.4.), each of which consist s of some customers depending on the capacity of the link and the codec type (Table 6-1). Also node 5 for FTP traffic is assumed to terminate at node 6 with a single shared 1 Mbps bottleneck link in the middle (link 7-8).

For the first scenario, an FTP source over TCP connection has been chosen between nodes 5 and 6 with 500 Kbit per second capacities. Furthermore, we use an exponential traffic generator (which generates traffic according to an exponential On/Off distribution) over a RTP connection for the voice links. Packets are sent at a fixed rate during On periods, and no packets are sent during Off periods. Both On and Off periods are exponentially distributed with the following parameters:

- The constant size of the packets is 100 Bytes.
- The average‘ on’ time is 500 msec.
- The average‘ off’ time is 500 msec.
- The sending rate during ‘on’ times is 300 Kb/s.

We choose 20 bytes voice frame per 20 msec and it is encapsulated in RTP which has a 12 bytes header, UDP 8 bytes and IP 20 bytes in sequence. The required capacity is 23.43 Kbit per second and its packet size is 480 bits. However, 13 voice sources can be

delivered with 300 Kbit per second link capacity. The voice sources which use voice activity detection technology, do not employ every voice frame interval (the exponential source can simulate this situation), so 22 voice sources can mostly be delivered with voice activity detection. Table 6-2 shows the above parameter for other annexes of the G.729. Also there is drop tail queue management with 15 maximum buffer size in the link between nodes 7 and 8.

Coder	G.729	Annex D	Annex E
Frame length	20 bytes	16 bytes	29.5 bytes
Frame size	20 msec	20msec	20msec
Silence Detection	Enabled	Enabled	Enabled
RTP/UDP/IP Header	40 bytes	40 bytes	40 bytes
Packet size	60 bytes	56 bytes	69.5 bytes
Required capacity	23.43 Kbit/sec	21.9 Kbit/sec	27.15 Kbit/sec

Table 6-2: Packet parameters for G.729 series.

The simulation was run for the BE only (Table 6-3 depicts the transmitted packets and lost packets in BE service) and with differentiated services using priority queuing with three queues; two for voice and the other for BE. The aggregate traffic produced by all VoIP sources is restricted to be 70 percent of the bottleneck link capacity of 1.0 Mbps. The remaining capacity is allocated to the BE traffic.

Source-Destination	No. of Transmitted Packets	No. of Dropped Packets	Retransmitted Packets
0 – 4	232	26	0
1 – 4	814	51	0
2 – 4	572	70	0
3 – 4	289	18	0
5 – 6	179	4	4

Table 6-3: Simulation result for the best-effort.

The BE traffic is handled at the edge router using a random early detection (RED) queue with the RED parameters min_{th} =10, max_{th} =15, w_q =0.002 and max_p =0.02 [FJ93] . We have considered homogeneous flows where voice traffic aggregate consists of packets generated by sources that use the same codec algorithm. Here we can study the traffic generated by various codecs listed in Table 6-2. The voice sources are modeled as exponential bit rate (EXP) sources. In all the experiments, if EF is used to transport the voice traffic aggregates, then the EF traffic flows are allocated the subscribed rate of the EF class. The instantaneous capacity is determined in part by the structural characteristics of the connection, such as its bottlenecks, and in part by the traffic dynamics.

As discussed in Section 5.2, the DS entails that the edge router encodes class information into the header of the IP packet. The core router simply classifies incoming packets based on the class information. Each core router maintains simple queue statistics for QoS in the voice class and BE class. The throughput of traffic classes will be constrained during the congestion periods by the maximum number of packets limitation for each traffic class. As a result, each class gets only a limited fraction of the link bandwidth. Whenever a QoS-voice packet arrives, the statistics are updated and compared against a queue occupancy threshold that indicates the maximum number of QoS-voice packets in the queue. If the updated value exceeds the threshold, the incoming

QoS-UDP packet is dropped. Otherwise, the packet is queued up. Network state results corresponding to two capacity settings for the bottleneck link are given in Table 6-4 and 6-5. In Figure 6-3 the delay that occurs for an FTP service or a BE traffic is specifically shown. As can be seen from the results (Tables 6-4 and 6-5), for Diffserv architecture the delay increase for the voice traffic is negligible in comparison with that for the FTP packets which increased dramatically.

Source-Destination	No. of Transmitted Packets	No. of Dropped Packets	Mean Delay (msec)	Mean Delay without DS
0 – 4	232	0	18.9	15
1 – 4	814	0	20.6	15
2 – 4	572	0	18.8	15
3 – 4	289	0	17.8	15
5 – 6	179	0	328.8	26.6

Table 6-4: Simulation result with DS for 1 Mbit per second bottleneck capacity.

Source-Destination	No. of Transmitted Packets	No. of Dropped Packets	Mean Delay (msec)	Mean Delay without DS
0 – 4	232	0	18.9	15.2
1 – 4	814	0	28.7	15.2
2 – 4	572	0	18.9	15.2
3 – 4	289	0	20.2	15.2
5 – 6	179	0	435.6	28.4

Table 6-5: Simulation result with DS for 800 Kbit per second bottleneck capacity.

6.3 De-jitter Aware Adaptive VCBRBC

Packet switching technology is now used to carry traffic of all types in a uniform format as a stream of packets, each containing a header with networking information and a payload of bytes of data. The overall objective in transmission over IP networks is to

find an array of technical solutions to guarantee the desired level of the QoS under most network conditions.

To add redundancy to the voice stream at the sender node is widely accepted as a means to reduce the effective packet loss observed by the receiver. Redundant information is transmitted along with the original information in VCBR, so that the lost original data can be recovered, at least in part, from the redundant information. On the other hand, VCBR methods have been developed to compensate for packet loss only, and cannot mitigate the delay jitter in the network. Moreover, the end-to-end delay is increased in most cases by applying VCBR schemes.

In VCBRBC (discussed in Section 4.2.2), a redundant voice stream is sent over an independent path of largely uncorrelated loss and delay characteristics. As a result, the probability of disturbance, such as packet erasure or excessive delay, which can impact all channels at the same time, will be small. Latency, also can be reduced in this approach by playing out the voice description through the path with lower delay, without any excessive computational complexity compared with multi-stream voice transmission over multiple paths [AS04].

Our proposed redundancy control algorithm is summarized in Figure 6-2. The delay budget estimated in accordance with the end-to-end delay concept, and initial and the final play-out delay are obtained using the distribution function of the queue waiting time according to the Pareto/M/1 queuing model.

However, at the beginning of the congestion period the network is not mainly affected; and according to our queue model, congestion effects can be mitigated by changing the playout delay in response to the network conditions. The optimum initial playout delay

for the de-jitter buffer is equal to the total variable delay in the connection (D_Q in the queue model).

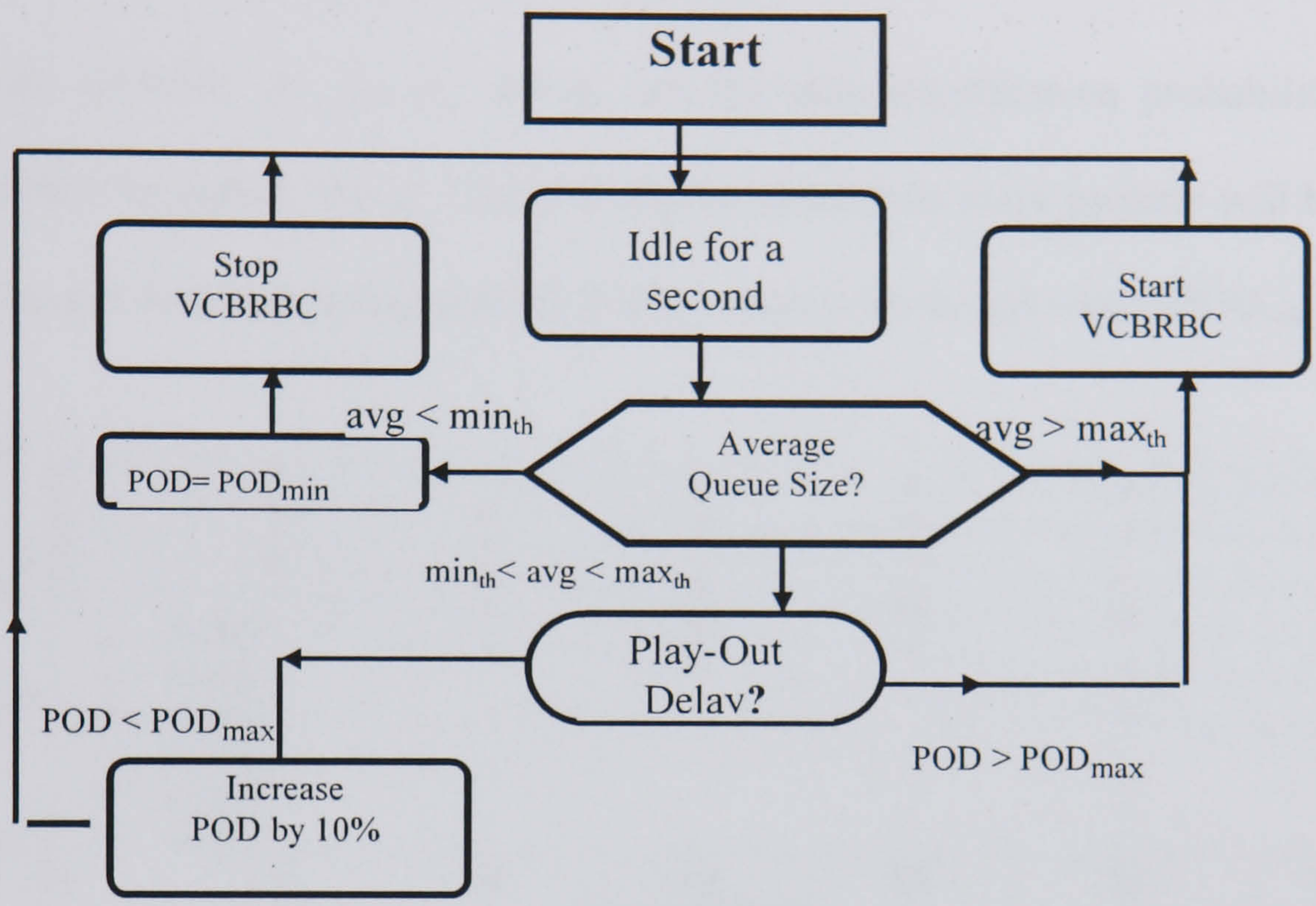


Figure 6- 2: Adaptive VCBRBC Algorithm.

The maximum depth of the buffer before it overflows is normally set to 1.5 or 2.0 times the initial playout delay. The contribution of the de-jitter buffer in the delay time comprises of the initial playout delay of the de-jitter buffer and the actual time interval for which the first packet is held in the network buffers. The late loss probability as a function of D_{Qmax} is shown in Figure 6-3 for different network conditions. By increasing the de-jitter buffer or playout delay (D_{Qmax}) the late loss probability decreases.

The network congestion and traffic load level are detected using the average queue size (*avg*) in the RED algorithm. The back up channel, and consequently the redundancy, will be employed when the *avg* exceeds max_{th} or in the cases where the play-out delay is set to its maximum value (POD_{min}). The loss probability of the network returns back into the

acceptable range. In the VCBRBC method, the loss probability of the overall system is $P_{l1}.P_{l2}$, in which $P_{l1} = \frac{p_1}{p_1 + q_1}$ and $P_{l2} = \frac{p_2}{p_2 + q_2}$ are the loss rates for path 1 and path 2 respectively [AS04]. p_1, q_1, p_2, and q_2, are the state transmission probabilities in the Gilbert model for path 1 and 2. Transmitting the redundant voice packets will be stopped when the *avg* is less than min_{th}, and the POD is reset to its initial value (POD_{min}).

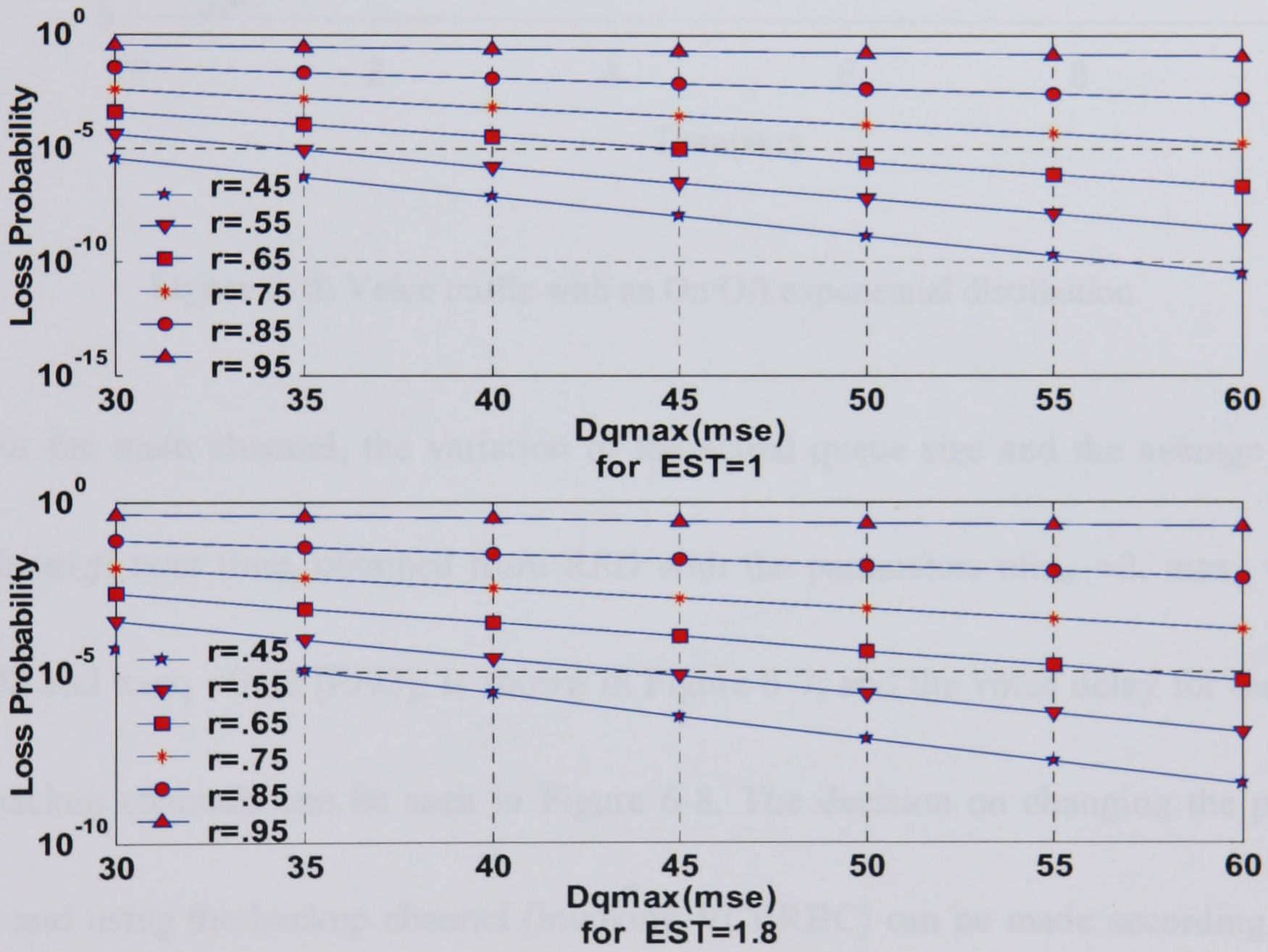


Figure 6- 3: Depicts loss probability versus Dqmax for different network conditions and two distinctive values of expected service time (EST).

6.3.1 Simulation Result

In this section the simulation and numerical results are given to demonstrate the performance of our adaptive VCBRBC algorithm. In the simulations we have considered a linear topology for the main and the backup paths depicted in Figure 6-4. Each path has

PAGE NUMBERING AS ORIGINAL

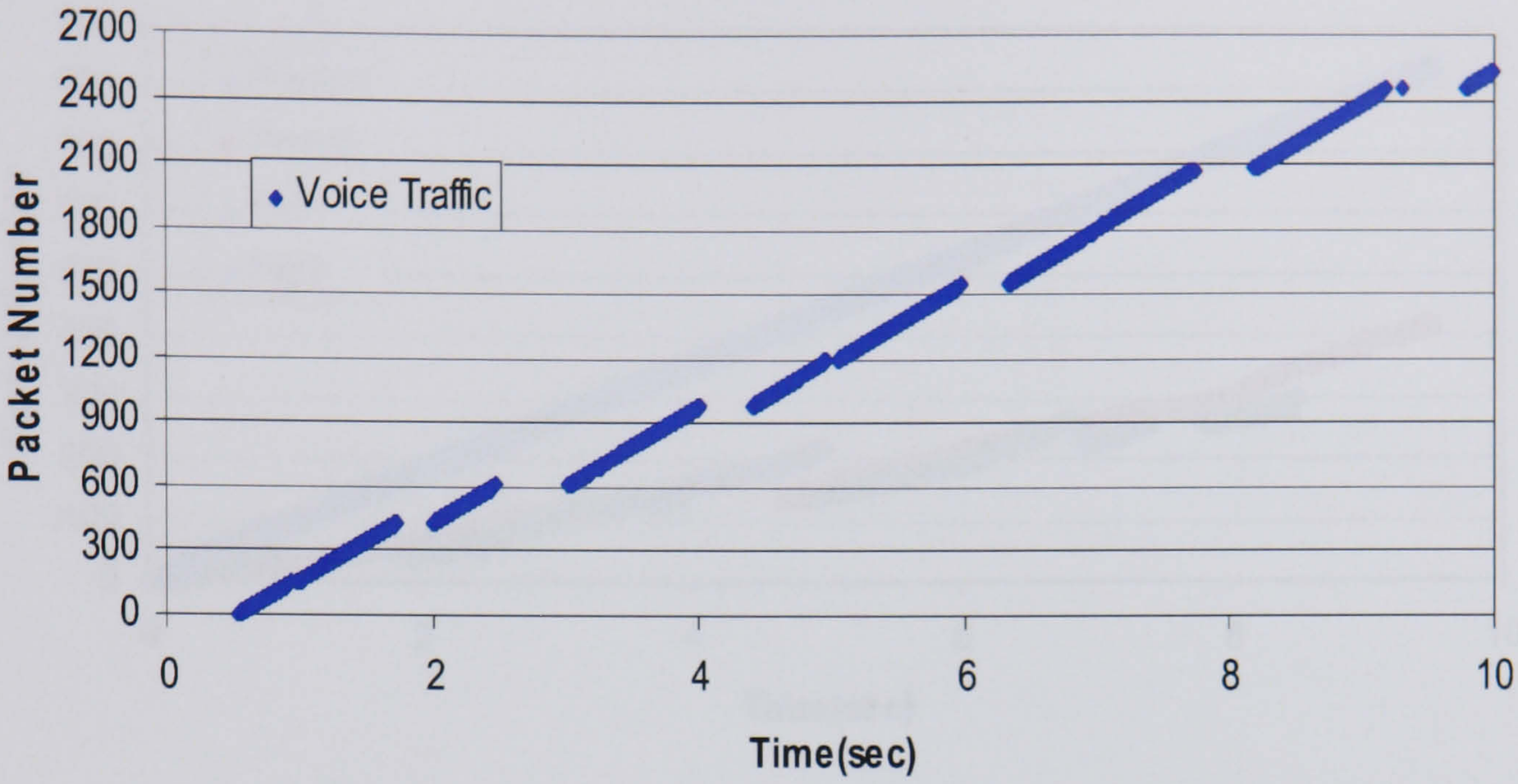
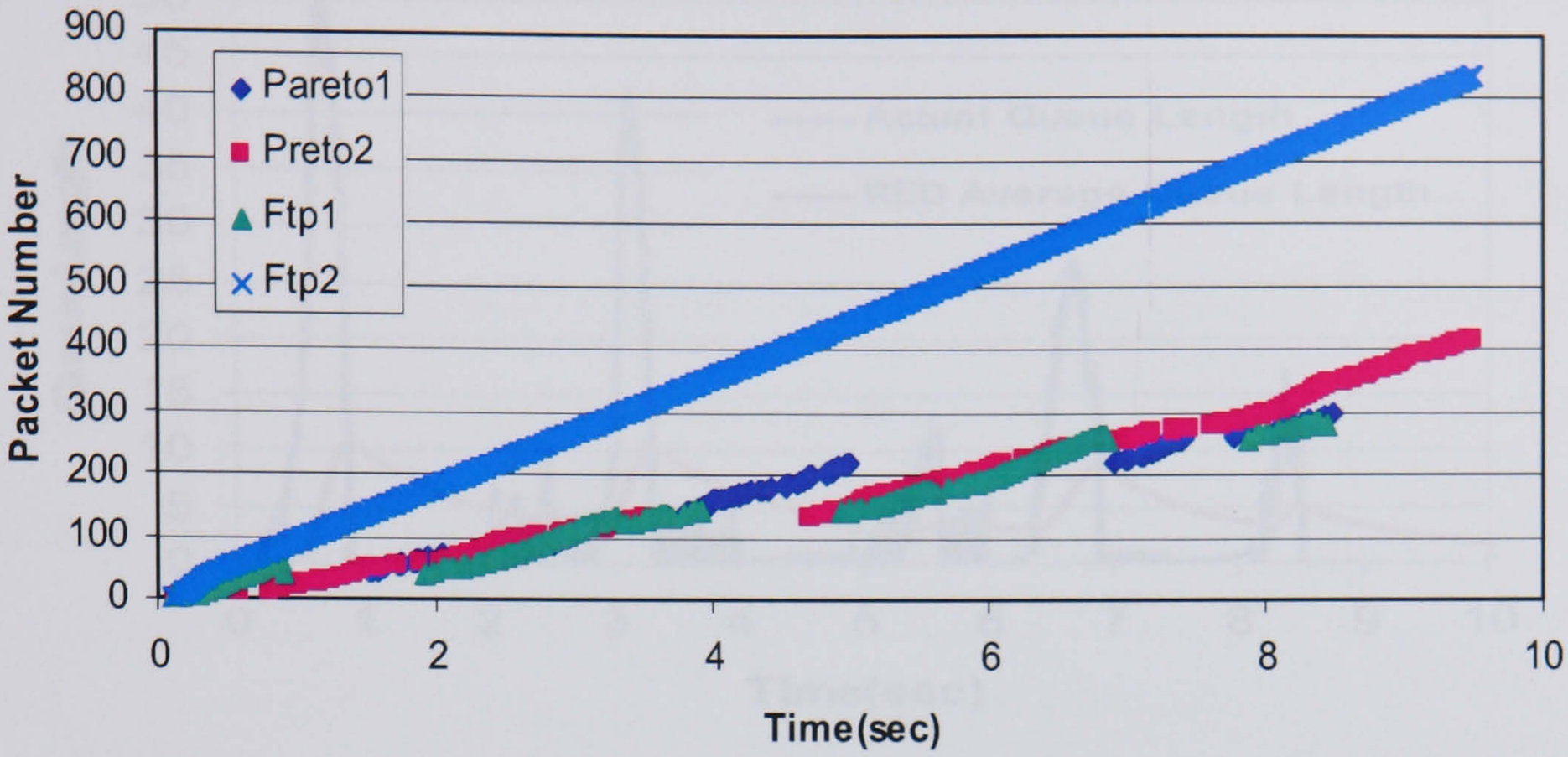
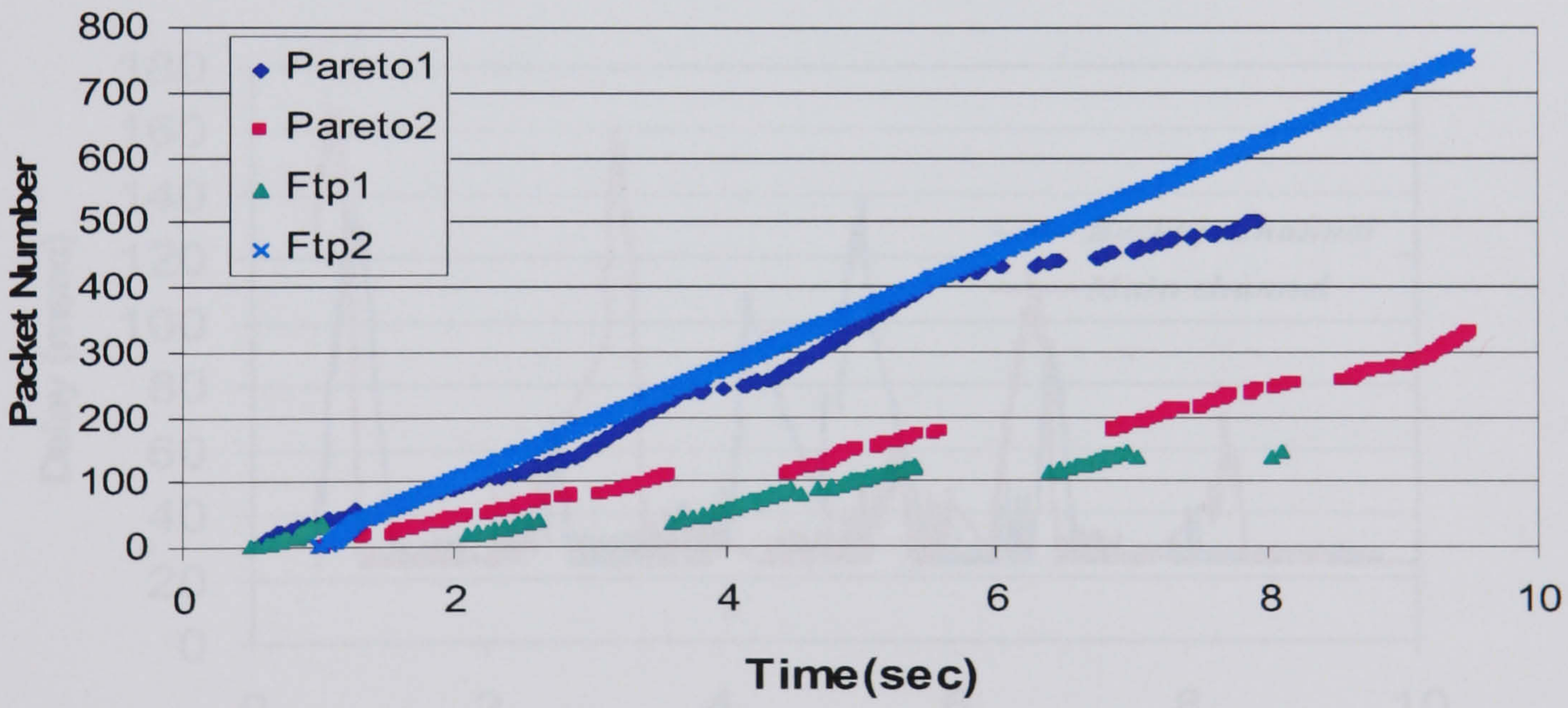


Figure 6- 5: Voice traffic with an On/Off exponential distribution.

For the main channel, the variation of the actual queue size and the average queue length (***avg***) over time, obtained from RED with the parameters min_{th} =2, max_{th} =5, w_q =0.002 and max_p =0.02 [FJ93], is shown in Figure 6-7; and the voice delay for the main and backup channels can be seen in Figure 6-8. The decision on changing the playout delay and using the backup channel (invoking VCBRBC) can be made according to the average queue size and the packet voice delay.



a) Data traffic for main path.



b) Data traffic for backup path.

Figure 6- 6: Data traffic for main and backup path. Ftp1and Ftp2 are two FTP sources of traffic, and Pa1 and Pa2 are two Pareto-distributed sources of traffic.

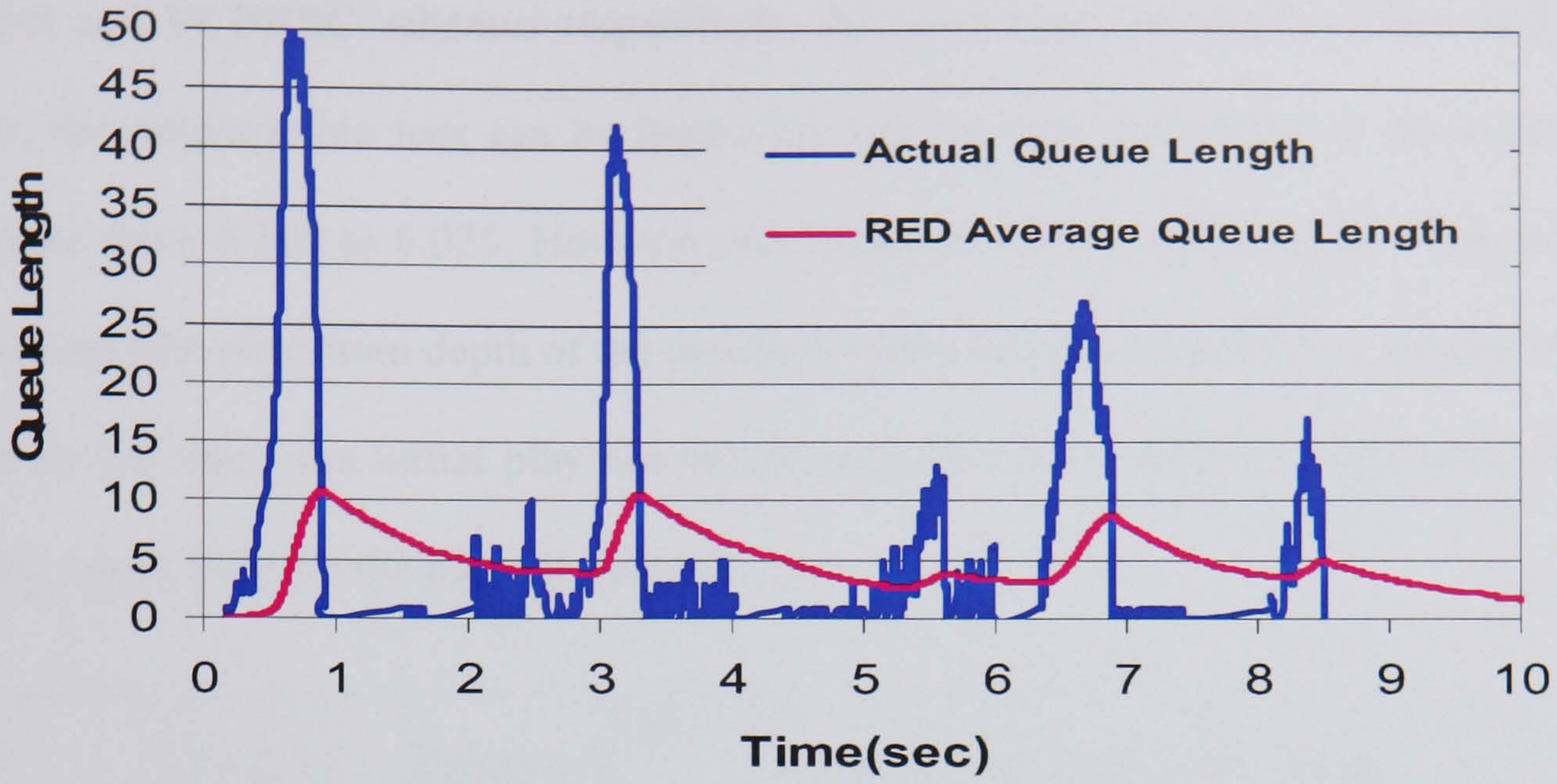


Figure 6- 7: Main path average (RED) and actual queue length.

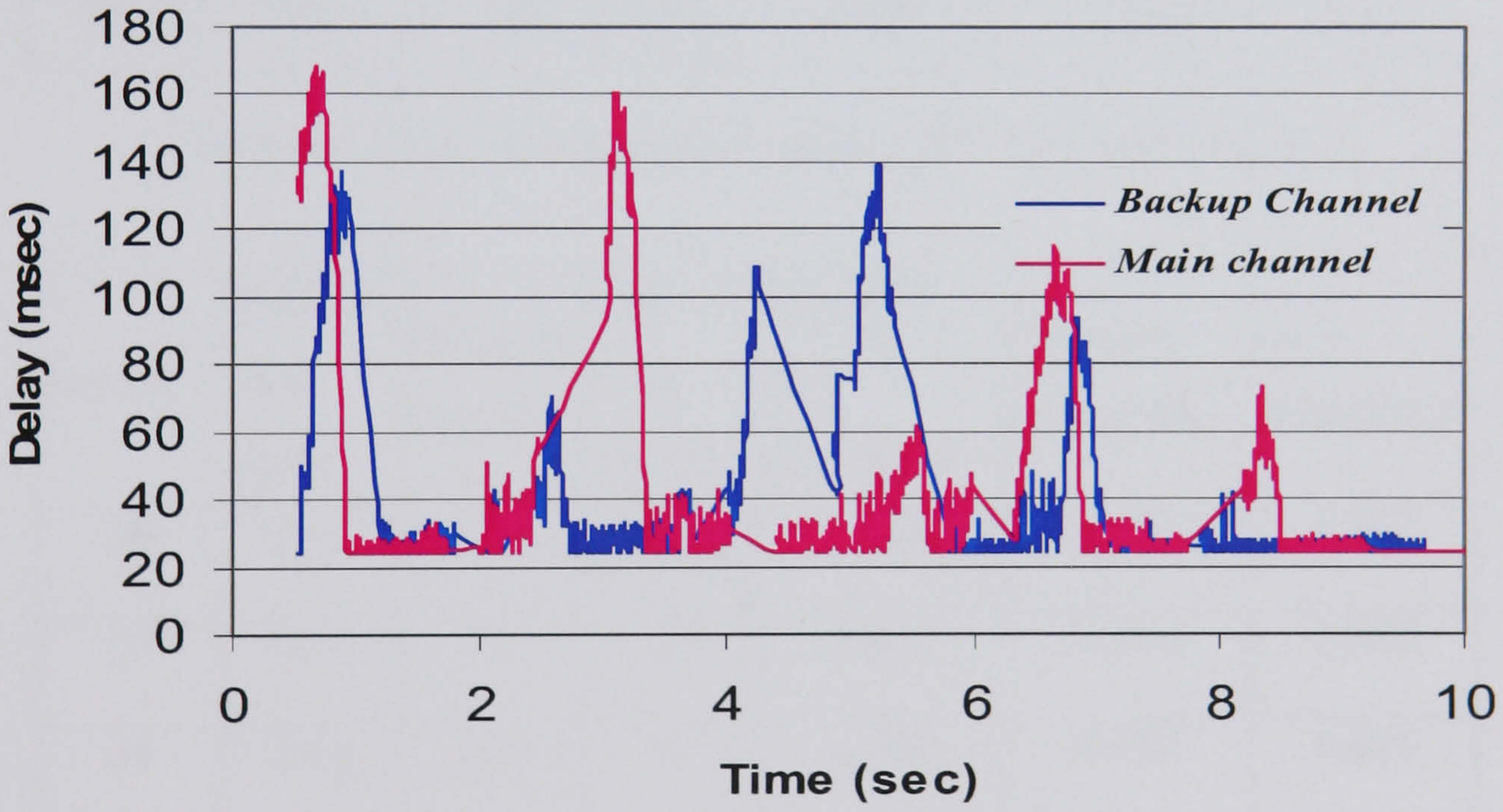


Figure 6- 8: Voice packet delay for main and backup paths over time.

The NS2 (network simulator) was used to obtain the numerical values for the transmitted, dropped, and late loss packet for different network conditions. The results are given in Tables 6-6 and 6-7 for various playout delay values, D_{Qmax}, and for single

channel and VCBRBC schemes respectively. By increasing the length of the de-jitter buffer, the network late loss can be improved, and the loss probability of the network decreases from 0.106 to 0.025. However, the variation of the de-jitter buffer length has restrictions (the maximum depth of the practical buffer before it overflows is normally set to 1.5 or 2.0 times the initial play out delay), and the ITU-T delay recommendation is another upper limit for the playout delay (D_{Qmax}).

Dqmax	Total of sent packet	Dropped Packet	Late Loss packet	Dropped Loss probability	Late Loss Probability	Loss Probability
40	2532	220	255	0.087	0.101	0.188
50	2532	220	228	0.087	0.090	0.177
60	2532	220	192	0.087	0.076	0.163

Table 6-6: Main path condition for different de-jitter values (D_{Qmax})

Dqmax	Total of sent packet	Dropped Packet	Late Loss packet	Dropped Loss probability	Late Loss Probability	Loss Probability
40	2532	42	181	0.016	0.071	0.087
50	2532	42	124	0.016	0.049	0.065
60	2532	42	74	0.016	0.029	0.045

Table 6-7: Adaptive VCBRBC network condition for different de-jitter values (D_{Qmax})

According to the adaptive algorithm (Figure 6-3), and using Figure 6-9 , we can obtain the time periods over which the backup channel needs to be employed in order to

decrease the loss probability so that *avg* becomes smaller than 5. For instant at t=0.635 seconds, the backup channel should open according to the de-jitter limit; and should open and then close at t=2.004 seconds due to the decreasing of packet delay and average queue size in the main channel. Figure 6-9 and Table 6-7 show the improvement of the adaptive network condition compared with the main channel. As one can see there is a 50% decrease in overall loss probability, and better voice delay condition in each path.

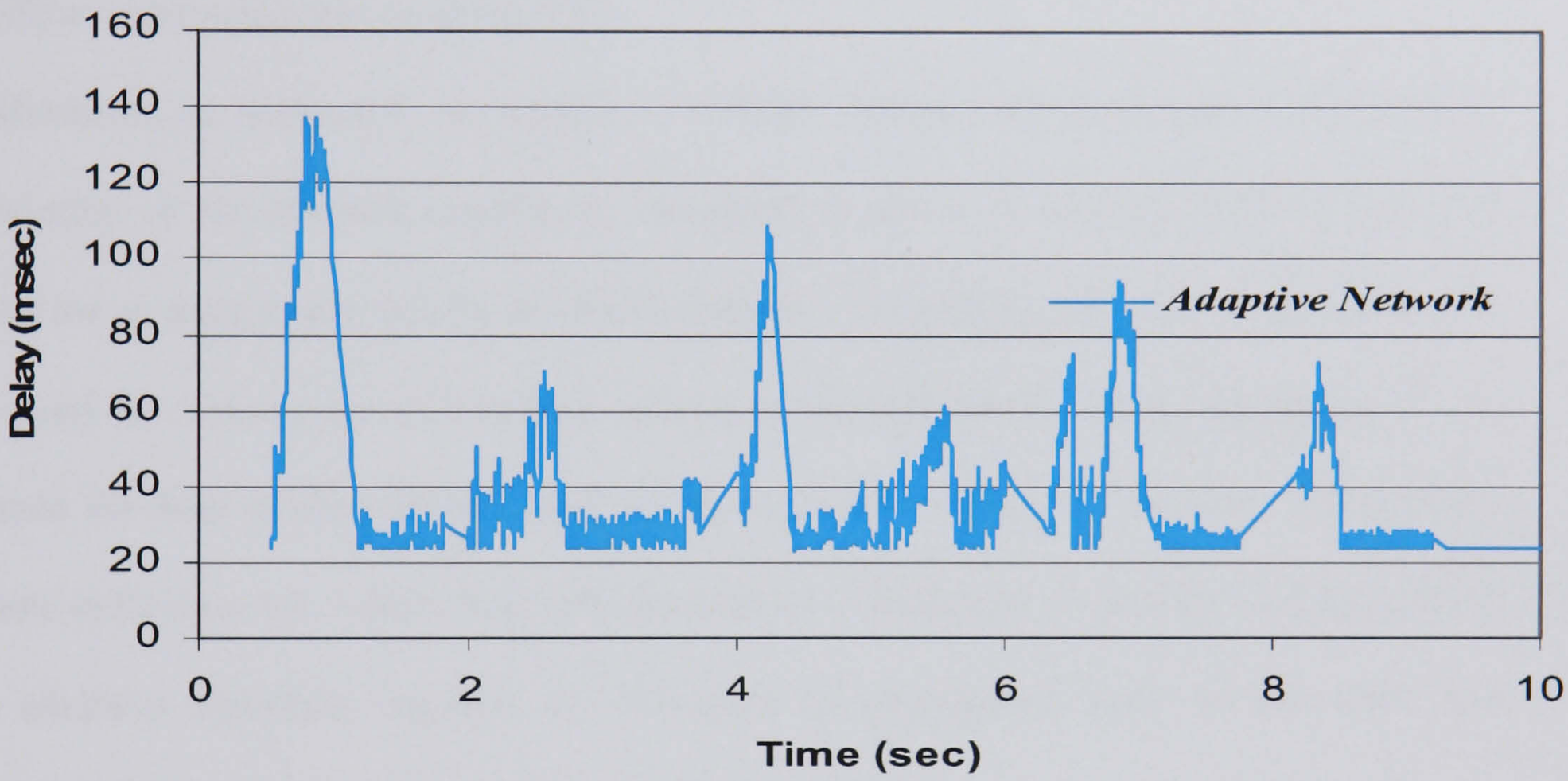


Figure 6- 9: Voice packet delay for adaptive network.

Although the percentage of the backup usage (43% in this simulation), entirely depends on the main path condition and QoS requirement, it is so clear that the adaptive system regardless to the prediction of the network condition can guarantee an efficient use of the network resources and better QoS than the plain applications.

6.4 Summary

In this chapter, we focused our attention on providing adaptability to IP telephony applications through variable bit-rate coding algorithms (in Section6.2) , and used a selective dropping mechanism to make the deadline misses distributed evenly during congestion. In Section6.2.1, t he performance of the proposed model was numerically evaluated by comparing the performance of simple drop-tail (FIFO) queuing, and RED in a Diffserv enabled network using NS2.

Moreover, in Section6.3 an adaptive VCBRBC scheme was proposed on the basis of estimation of the network conditions; measured in terms of de-jitter buffer length and using the average queue size in the RED method, to control the amount of redundancy or the need for making use of a backup channel, more efficiently. These algorithms aim to control the load of the network; and use the network resources efficiently. Simulation results in Section 6.3.1 show that 50% decrease in overall loss probability and the gain of the adaptive algorithm depends on difference in propagation delay of the main and backup paths, and the condition of the network traffic.

Chapter 7- Conclusion and Future Work

An acceleration of the merging of circuit-switched traffic (voice and fax) with packet-switching technologies could be seen in the future of telecommunication media. Integrated services, once envisioned in the circuit-switched world, will come to fruition in the packet-switched world. It may be that the IP telephony of today will become the unplanned source of the ultimate integration of voice and data. The proponents of packet switching should not beam too brightly, for the flexibility of the packet world is a two edged sword. VoIP refers to real-time delivery of voice packet across networks using the Internet protocol. VoIP's appeal is based on its capability to facilitate voice and data convergence at an application layer. One of the typical problems with the implementation of packet voice over IP is the difficulty of QoS guarantee. Voice quality is affected by available bandwidth, end-to-end delay, delay variation, acceptable error or loss rate without retransmission, and codec (coder/decoder) quality.

7.1 Conclusion

This research effort developed and evaluated two novel application sender base error correction schemes, to compensate for packet loss and mitigate the effect of delay jitter in the VoIP network. The results show that the bandwidth efficiency and the loss tolerance can improve by 17% and 50%, respectively, through increasing the required capacity for voice stream by less than the coding rate in the voice codec base redundancy scheme.

Also we have shown that over the Internet, the VCBRBC scheme can reduce the loss rate compared to the single path VCBR schemes. Uncorrelated delay variation and packet loss over different paths, is exploited in this method to retrieve the packet received from the backup channels (paths) in case of congestion in the main path. Picking the packet with minimum delay can reduce the average end-to-end delay whilst avoiding late loss or burst loss. The GML is invoked to assess the improvement in the packet loss probability in this new method compared with and without the codec-specific VCBR. Simulation results show that the gain of the backup channel depends on the difference in propagation delay of the multiple paths and the condition of the network traffic.

A new loss model for analysis of real-time voice transmission over IP networks is proposed, which can be used to analyze various techniques for enhancing the QoS in the IP telephony applications. We used the RED technique and de-jitter buffer respectively for estimation of the queuing delay and for tuning the packet loss in the receiver. Although we only model the voice traffic conveyed over the IP network, the effect of other traffics on the voice will be considered by using an appropriate queuing model.

The model predicts the network behavior with good accuracy provided that an efficient queue model and traffic estimation method are employed. The accuracy of the model is verified through simulation and analytical results for different traffic conditions, and it is shown that the model predicts the overall voice packet loss rate (late and dropped) over the Internet on the BE condition with good precision.

Adaptive error correction methods in VoIP applications have several appealing features, such as efficient use of the network resources, better QoS than the plain applications and availability of more resources for signaling and critical in-band flow

management sharing the same network facilities. Also, adaptive application in the Internet would reliably achieve the ability of making a trade-off between throughput, QoS, and utilization of the network.

In the first proposed adaptive algorithm, variable bit-rate voice coders adapt their rate to the time-varying network conditions by means of a control algorithm, whose aim is maximizing the utilization of the available bandwidth while reducing and preventing the occurrence of packet losses. By introducing the selective dropping model, we expect that consecutive packet loss could be avoided during the congestion. As it can be seen when Diffserv enabled, not only are there any packet losses in the congested network, but also we can increase the codec speed or instead reduce the bottleneck speed, which is done in the simulation. The performance of the proposed approach is evaluated in various scenarios which comprise a network dedicated to the exclusive use of adaptive voice sources and other data traffic such as FTP.

Secondly, a new adaptive VCBRBC algorithm for real-time voice transmission on the basis of estimation of the network condition using RED and de-jitter buffer length, has been proposed. It is shown that over the Internet on the BE condition, this new scheme can reduce the loss rate (late and dropped) and the network resource usage is improved compared to the usual VCBRBC scheme. Uncorrelated delay variation and packet loss over different paths is exploited in this method to retrieve the packet received from the backup channels (paths) in case of strong congestion in the main path. Using the network resources when it is absolutely necessary and picking the packet with minimum delay, can reduce the average end-to-end delay whilst avoiding late loss or burst loss.

Simulation results show that the gain of the adaptive algorithm depends on difference in propagation delay of the main and backup paths, and the condition of the network traffic.

7.2 Recommendation for Future Work

This research effort has extended the knowledge base of transporting real-time voice over general IP networks. A novel error correction scheme, VCBRBC, has been developed and that significantly improves the ability of IP networks to successfully transmit such data. Also a complete loss model, regarding voice traffic, for the IP networks has been introduced. While the improvements are noteworthy, extensions of this work may provide even more benefit. It is recommended that the following research areas be investigated.

1. Analyze the performance of VCBRBC using several real network paths with real data traffic instead of using some IP traffic models.
2. Investigate the dependency of the introduced model parameters to the network structure and the usage of the IP network.
3. According to the tremendous growth of Internet telephony, try to find a sender-based error correction and voice multiplexing scheme in a real network condition and not in a loss free network, regarding low end-to-end delay and high bandwidth efficiency, simultaneously.
4. Investigate performance improvement of the proposed error correction methods regarding the specified application and voice and video [TG04], conferencing; and try to adapt the method, VCBRBC and model, according to the application.

5. Verify the performance of the proposed error correction methods for real time multimedia traffic, voice and video, over IP networks.
6. Investigate the scalability of the proposed adaptive error correction schemes in the VoIP applications.
7. Investigate the effect of the multi-queue system instead of single-queue which we considered in spite of the packet size of the voice, (very small compared to the other data traffic) and which can mitigate the effect of a multi-queue system in our approach to define the model parameters.

Bibliography

[ABFRV02] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "Multiscale nature of network traffic," *IEEE Signal Processing Magazine*, Vol. 19, No. 3, pp.28-46,May 2002.

[AMS03] A. Asosheh, F. Marvast*i*, and M. Shikh-Bahae*i*, "Speech-Property-Based FEC algorithm for VoIP," *International Symposium on Telecommunication, IST2003*, pp.105-109, August 2003.

[AS03] A. Asosheh and M. R. Shikh-Bahaei, "Adaptive voice over IP network with selective dropping mechanism", *4th International Conference on 3G Mobile Communication Technologies*, London, pp. 404-408, June 2003.

[AS04] A. Asosheh, and M. Shikh-Bahaei, "Voice over adaptive IP network (VoAIP)," *IEEE Australian Telecommunications Network and Applications Conference (ATNAC)*, Sydney, pp. 15-21,December 2004.

[ASC04] A. Asosheh, M. Shikh-Bahaei, and J. A. Chambers, "QoS enhancement for VoIP using a new FEC scheme with backup channel," *IEICE Transaction On Communication,* Vol. E87-B, No. 10, pp. 3101-3106, October 2004.

[BB04] R.M. Bahati, M.A. Bauer, "Quality of service provisioning for VoIP applications with policy-enabled differentiated services," *IEEE/IFIP Network Operations and Management Symposium,* Vol. 1, pp.335-348, April 2004.

[BBKT97] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. Tripathi, "Using channel state dependent packet scheduling to improve TCP throughput over wireless LANs," *Wireless Networks*, Vol. 3, No. 1, pp. 91-102, March 1997.

[BCDM01] A. Barberies, C. Casetti, J.C. De martin, and M. Meo, "A simulation study of adaptive voice communications on IP networks," *Elsevier Computer Communications,* Vol 24, No.9, pp. 757-767, May 2001.

[BCR01] F. Beritelli, S. Casale, and G. Ruggeri, "Performance comparison between VBR speech coders for adaptive VoIP applications," *IEEE Communications Letters,* Vol. 5, No. 10, pp. 423-4256, October 2001.

[BF01] C.C. Bear*d*, and V.S. Fros*t*, "Prioritized resource allocation for stressed networks," *IEEE/ACM Transactions on Networking*, Vol. 9, No. 5, pp. 618 633, October 2001.

[BG96] J-C. Bolot and A. Garcia, "Control mechanisms for packet audio in the Internet," *Proceedings IEEE INFOCOM*, pp. 232 - 239, March 1996.

[BGMT98] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, "Queuing Networks and Markov Chains," *John Wiley & Sons, Inc*, 1998.

[BPT99] J-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-Based error control for Internet telephony," *Proceedings IEEE INFOCOM*, Vol. 3, pp. 1453-1460, March 1999.

[Bra69] P. T. Brady, "A model for generating on/off speech patterns in two-way conversations," *Bell Systems Technical Journal*, 2445-2472, September

1969.

[BRCDS94] Braden, R., Clark, D. and S. Shenker , "Integrated services in the Internet architecture: An overview," *RFC 1633*, July 1994.

[Bro00] K.Brown, "The RTCP gateway: scaling real-time control bandwidth for wireless networks," *ElsevierComputer Communications* , Vol. 23, No.14-15, pp.1470-1483, August 2000.

[BWW99] F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB," RFC 2597*IETF Draft* , January 1999.

[Cin75] E. Cinlar, "Introduction to Stochastic Processes," *Prentice-Hall*, 1975.

[DMM88] S. Dravida, M. J. Master, and C. H. Morton, "A method to analyze performance of digital connections," *IEEE Transactions on Communications*, Vol. 36, No.3, pp. 298 305, March 1988.

[DS99] C. Dovrolis and D. Stiliadis, "Relative differentiated service in the Internet: Issues and mechanisms," *In Proc. ACM SIGMETRICS,* pp.204-205, May 1999.

[DSR02] C. Dovrolis, D. Stiliadis, and P. Ramanathan, "Proportional differentiated services: Delay differentiation and packet scheduling," in IEEE/*ACM Transaction on Networking*, Vol. 10, No. 1, pp. 12-26, February 2002.

[EH99] E. Ekudden, and R. Hangent, "The adaptive multi-rate speech coder," *IEEE Speech Coding Proceedings,* pp.117-119 June 1999

[FH99] M. J. Fischer, and C. M. Harris, "A method for analyzing congestion in Pareto and related queues," *The Telecommunications Review,* Mitretek

Systems, pp.15-27, 1999.

[Fin02] Victoria Fineberg, "A practical architecture for implementing end-to-end QoS in an IP network," *IEEE Communications Magazine,* Vol. 1, No.1, pp.122-130, January 2002.

[FJ93] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, Vol. 1, No.4, pp.397-413, August 1993.

[FM94] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, Vol. 32, No.3, pp. 70-81, March 1994.

[FNYF03] H. Furuya, S. Nomoto, H. Yamada, N. Fukumoto, F. Sugaya, "Experimental investigation of the relationship between IP network performances and speech quality of VoIP," *IEEE/ ICT Telecom-munications,* Vol. 1, pp. 543 552, March 2003.

[FP01] S. Floyd, and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, Vol. 9, No. 4, pp. 392-403, August 2001

[FV05] K. Fall, and K. Varadhan, "The NS manual: Formerly NS notes and documentation," www.isi.edu/nsnam/ns/doc/ns_doc.pdf , November 2005.

[GB97] S. Giordano, J.Y. Le Boudec, "VBR over VBR: The homogeneous, loss free Case," *Proceeding IEEE Computer and Communications Societies*

INFOCOM, Vol. 1, pp. 168-176, March 1997.

[Gil00] R.H. Gilito, "Advanced services architectures for internet telephony: A critical overview," *IEEE Transactions on Networking*, Vol. 14, No.4, pp. 38-44,July -August 2000.

[Gil60] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, Vol.39 , No. 2, pp.1253-1265, 1960.

[GPM03] K. P. Gummadi, M.J. Pradeep, C.S.R. Murthy, "An efficient primary-segmented backup scheme for dependable real-time communication in multihop networks," *IEEE/ACM Transactions on Networking,* Vol. 11, No. 1, pp. 81-94, February 2003.

[Gro02] D. Grossman, "New terminology and clarification for Diffserv," *RFC 3260, IETF,* April 2002.

[Hao02] Fang Hao, "QoS routing for any cast communications; motivation and an architecture for DiffServ networks," *IEEE Communications Magazine,* Vol. 40, No. 6, pp. 48-56, June 2002.

[HMP00] T. Harbaum, D. Meter and M. Prinke, "Hardware support for RSVP capable routing," *Proceedings of the IEEE Conference on High Performance Switching and Routing*, pp. 241-249, June 2000.

[HNA00] M. Hassan, A. Nayandoro, and M. Atiquzzaman, "Internet telephony: Services, technical challenges, and products," *IEEE Communication Magazine*, Vol. 38, No. 4, pp. 96–103, April 2000.

[HRR05] F. Hammer, P. Reichl, A. Raake, "The well-tempered conversation:

interactivity, delay and perceptual VoIP quality," *IEEE International Conference on Communications ICC,* Vol 1, pp. 244 249, May 2005.

[HS98] S. Han, and K.G. Shin, "A Primary-Backup channel approach to dependable real-time communication in multi-hop networks," *IEEE Transaction on Computer,* Vol. 47, No. 1, pp. 46 61, January 1998.

[HTW04] E. Haddadeh, R. Taylor, G.A. Watts, "Towards scalable end-to-end QoS provision for VoIP applications," *IEE Telecommunications Quality of Services: The Business of Success,* pp. 132 135, March 2004.

[HWJ05] J. Han, D. Watson, F. Jahanian, "Topology aware overlay networks," *Proceeding IEEE INFOCOM,* Vol. 4, pp. 2554 – 2565, March 2005.

[IETF01] Internet Engineering Task Force "SIP: Session Initiation Protocol," *IETF Internet Draft*, May 2001.

[ITU00] International Telecommunications Union, "One-way transmission time," *ITU-T Recommendation G.114*, May 2000.

[ITU96] International Telecommunications Union, "Subjective performance assessment of telephone-band and wideband digital codecs," *ITU-T Recommendation*, P.830, 1996.

[ITU99] ITU-T Recommendation G.729, "General aspects of digital transmission systems," 1999.

[ITUG00] ITU-T Recommendation G.729 – Annex I, "Transmission systems and media, digital systems and networks," February 2000.

[JA05] Y. Jung, J.W. Atwood, "Switching between fixed and call-adaptive

playout: A per-call playout algorithm," *IEEE Internet Computing,* Vol. 9, No. 4, pp. 22-27, July 2005.

[JC81] N. Jayant and S. W. Christensen, "Effects of packet losses in waveform coded speech and improvements due to an Odd-Even sample-interpolation procedure," *IEEE Transactions on Communications*, Vol. 29, No. 2, pp. 101 - 109, February 1981.

[JCG04] H. James, B. Chen, L. Garrison, "Implementing VoIP: A voice transmission performance progress report," *IEEE Communications Magazine,* Vol. 42, No. 7, pp. 36-41, July 2004.

[JNP99] V. Jacobson, K. Nichols, and K. Poduri, "An expedited forwarding PHB," *RFC 2598 IETF Draft*, June1999.

[JR86] R. Jain and S. A. Routhier, "Packet trains-measurements and a new model for computer network traffic," *IEEE Journal on Selected Areas in Communications,* Vol. 4, No. 6, pp. 986-995, September 1986.

[KB96] E. Knightly, and H-BIND, "A new approach to providing statistical performance guarantees to VBR traffic," *Proceedings IEEE INFO-COM San Francisco*, Vol. 3, pp. 1091-1099, March 1996.

[Kle76] L. Kleinrock, "Queuing Systems," Volume II, *John Wiley and Sons*, 1976.

[KM01] K.Krishna, V.L.N.Murty , "Vector quantization of excitation gains in speech coding ," *Elsevier Signal processing,* Vol. 81, No. 1, pp. 203-209, January 2001.

[KT01] M. J. Karam, and, F. A. Tobagi, "Analysis of the delay and jitter of voice

traffic over Internet," *Proceedings IEEE INFOCOM*, Vol. 2, pp. 824-833, April 2001.

[KY05] Y. H. Korhonen, and J. YeWang, "Optimization of source and channel coding for voice over IP," *IEEE International Conference on Multimedia and Expo,* pp. 173 176, July 2005.

[LI04] S.' Lingfen, E. Ifeachor, "New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks," *IEEE International Conference on Communication,* Vol. 3, pp. 1478– 1483, June 2004.

[LPD02] S.H. Low, F. Paganini, and J.C. Doyle , "Internet congestion control," *IEEE Control System Magazine,* Vol. 22, No. 1, pp. 28-43, February 2002.

[LS95] S. Ben Slimane and T. Le-Ngoc, "A doubly stochastic Poisson model for self similar traffic" *Proceeding IEEE International Conference on Communications*-ICC, Vol. 1, pp. 456-460 1995.

[LSG01] Y. J. Liang, E. G. Steinbach, and B. Girod, "Multi-stream voice over IP using packet path diversity," *IEEE Multimedia Signal Processing*, pp. 555 -560, October 2001.

[LTWW94] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 1-15, February 1994.

[Lu00] G. Lu, "Issues and technologies for supporting multimedia

communications over the Internet," *Elsevier Computer Communications,* Vol. 23, No. 14-15, pp. 1323-1335, August 2000.

[MBK02] V. Matié, A. Ba, and M. Kos, "Voice traffic performance measurement in packet networks," *24th International Conference on Information Technology Interfaces IT*, pp. 499-504, June 2002.

[Mup00] J. K. Muppala, "VoIP performance on differentiated services enabled network," *IEEE International Conference on Networks,* pp.419-423, June 2000.

[NCP99] A. Neogi, T. Chiuch, and P. Stirpe, "Performance analysis of an RSVP-capable router," *IEEE Network,* Vol. 13, No. 5, pp.56-63, October 1999.

[Nol95] P. Noll, "Digital Audio Coding for Visual Communications," *Proceedings IEEE,* Vol. 83, No. 6, pp. 923-945, June 1995

[Nor95] I. Norros, "The management of large flows of connectionless traffic on the basis of self-similar modeling," *Proceedings IEEE International Conference on Communication ICC*, Vol. 1, pp. 451-455, 1995.

[OC04] E. H. Orallo, and J. V. Carbo, "In advance activation of backup channels for real-time transmission," *Dependable Systems and Networks Conferences*, pp. 555– 560, July 2004.

[PB86] T. Pratt and C. W. Bostian, "Satellite Communications," *John Wiley and Sons*, 1986.

[PF95] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp.

226 244, June 1995.

[PH01] G. Priggouris, and S. Hadjiefthymiades, "GPRS + IntServ / RSVP: an integrated architecture," *Elsevier Computer Networks*, Vol. 37, No. 5. pp. 617 630, November 2001.

[PHH98] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Transaction on Networks*, Vol. 12, No. 5, pp. 40 - 48, September-October 1998.

[PJAM04] S. Praestholm, S.S. Jensen, S.V. Andersen, M.N. Murthi, "On packet loss concealment artifacts and their implications for packet labeling in voice over IP," *IEEE International Conference on Multimedia and Expo,* Vol. 3, pp. 1667 1670, June 2004.

[Pru95] P. Pruthi, "Heavy-Tailed On/Off source behavior and self-similar traffic," *IEEE International Conference on Communications*, pp. 445-450, February 1995.

[RAHB03] L. Roychoudhuri, E. Al-Shaer, H. Hamed, G.B. Brewster, "Audio transmission over the Internet: experiments and observations," *IEEE International Conference on Communication ICC,* Vol. 1, pp. 552-556, May 2003.

[RB01] G. Ruggeri, F. Beritelli, "Hybrid multi-mode/multi-rate CS-ACELP speech coding for adaptive voice over IP," *Proceeding IEEE/ ICASSP,* Vol. 2, pp. 733 736, April 2001.

[Rez99] J. F. Rezende , "Assured service evaluation ," *Proceedings IEEE Global*

Communications Conference GlobCom, Vol. 1a, pp. 100-104, December 1999.

[Rin99] J. Rinde, "Telephony in the year 2005," *ElsevierComputer Networks* , Vol. 31, No.1, pp. 157 167, 1999.

[RLSG04] H. Ruibing, D. Lee, R.K. Sinha, N. Griffeth, "Integrated system interoperability testing with applications to VoIP," *IEEE/ACM Transactions on Networking,* Vol. 12, No. 5, pp. 823-836, October 2004.

[ROZY05] T. Ruixiong, Z. Oian, X. Zhe, X. Yongqiang, " Robust and efficient path diversity in application-layer multicast for video streaming," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 15, No. 8, pp. 961 – 972, August. 2005.

[RP00] H. Ren, and K. Park, "Toward a theory of differentiated services," *In Proceeding IEEE/IFIP IWQOS*, pp. 211-220, 2000.

[RR02] M. Reisslein, K.W. Ross, "A framework for guaranteeing statistical QoS," *IEEE ACM Transaction on Networking*, Vol. 10, No. 1, pp. 27-42, February 2002.

[RS04] P. Renjie, and J. Song, "Multi-path transmission based on overlay network," *Advanced Information Networking and Applications AINA,* Vol. 2, pp. 330 333, August 2004.

[San98] H. Sanneck, "Concealment of lost speech packets using adaptive packetization," *Proceedings IEEE Multimedia Systems,* pp. 140-149, June 1998.

[Sch88] M. Schwartz "Telecommunication Networks: Protocols, Modeling and Analysis" *Addison-Wesley,* 1988.

[Sch99] H. Schulzrinne, "The IETF Internet telephony Architecture and Protocols," *IEEE Network,* Vol. 13, No. 3, pp. 18-23, May/June 1999.

[SFC96] H. Schulzrinne, GMD Fokus and S. Casner 'RTP: A transport protocol for real-time applications' *RFC 1889, IETF,* January 1996.

[SKS01] J. Schmitt, M. Karsten, R. Steinmetz, "On the aggregation of deterministic service flows," *Elsevier Computer Networks,* Vol. 24, No.1, pp. 2-18, 2001.

[SLLY02] H. P. Sze, S.C. Liew, B. Lee, and D. C. S. Yip, "A multiplexing scheme for H.323 voice-over-IP applications," *IEEE, Journal of Selected Areas in Communication*, Vol. 20, No. 7, pp. 1360-1368,September 2002.

[SP00] T.L. Sheu, G.Y.Pao , "Design of a two-pass RSVP setup mechanism for integrated services," *IEEE International Conference on Communication ICC,* Vol. 2, pp. 718-722, June 2000.

[SRM97] K. Sundstrom, A. Rueda, R. D. Mcleod, "Internet telephony compression algorithms," *IEEE Conference of Communications Power and Computing,* pp.13 18, May 1997.

[SSR99] B. Subbiah, S. Sengodan and R. Rajahalme, "RTP payload multiplexing between IP telephony gateways," *IEEE Global Communications Conference GlobCom*, Vol. 2, pp. 1121-1127, December 1999.

[Sul01] A. Sulkin, "Next-Generation IP Phones Arriving," *Business Communications Review,* pp. 29-32,D ecember2001 .

[TG04] S. Tao, R. Guerin, "Application specific path switching: A case study for streaming video," *ACM International Conference on Multimedia,* pp 136-145, October 2004.

[Tho96] G.A. Thom, "H.323: The multimedia communication standard for local area networks," *IEEE Communications Magazine,* Vol. 34, No. 12, pp. 52-56,December 1996.

[THS03] X. Tang, J. Huang, G.B. Siew, "QoS provisioning using IPv6 flow label in the Internet," *International Conference on Information, Communications and Signal Processing,* Vol. 2, pp. 1253-1257, December 2003.

[TJ00] G. Thomsen and Y. Jani, "Internet telephony: Going like crazy," *IEEE Spectrum*, Vol. 37, No. 5, pp. 52–58, May 2000.

[Tog99] J. Toga, "ITU-T standardization activities for interactive multimedia communications on packet-based networks: H.323 and related recommendations," *Elsevier Computer Networks,* Vol 31, No. 3, pp. 205–223, February 1999.

[TXEG05] S. Tao, K. Xu, A. Estepa, T.F.L Gao, "Improving VoIP quality through path switching," *Proceedings IEEE Communications Societies,* Vol. 4, pp. 2268-2278, March 2005.

[TXXF04] S. Tao, K. Xu, Y. Xu, T. Fei, "Exploring the performance benefits of end-

to-end path switching," *IEEE International Conference on Network Protocols,* pp. 304 315, November 2004.

[VA99] S. Vutukury, L. Aceves, "A scalable architecture for providing deterministic guarantees," *Proceedings IEEE On Computer Communications and Networks,* pp. 91-96, October 1999.

[VZ95] M. A. Visser and M. El Zarki, "Voice and data transmission over an 802.11 wireless network" *In 6th IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications*, Vol. 2, pp. 648-652 September 1995.

[WCHLT04] P.C. Wang, C.T. Chan, S.C. Hu, C.L Lee, W.C Tseng, "High-speed packet classification for differentiated services in next-generation networks," *IEEE Transactions on Multimedia,* Vol. 6, No. 6, pp. 925-935, December 2004.

[WH01] C. Y. Wang, C. Hsu Y. Huang, "VORAL: A system for voice over IP routing in application layer," *7th IEEE symposium Real-Time Technology and Applications,* pp. 165 170, June 2001

[WM01] F. Wang, P. Mohapatra, "Using differentiated services to support internet telephony," *Elsevier computer communications,* Vol. 24, pp. 1846-1854 , December 2001.

[WP01] White Paper, "Audio codecs and cisco unity," *Cisco System Inc*2001.

[Wri01] David J. Wright, "Voice over Packet Networks," *John Wiley & sons,* 2001.

[WS03] K. Wenyu Jiang, and H. Schulzrinne, "QoS evaluation of VoIP end-points," *IEEE International Conference on Communication ICC,* Vol. 3. pp. 1917– 1921, May 2003.

[WTS97] W. Willinger, M.S. Taqqu, and R. Sherman, "Self-Similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, February 1997.

[XXL96] Y. Féng, W. Xu, T.X. Lu, "About multimedia communication standardization architecture," *Proceeding IEEE International Conference on Signal Processing ICSP'* 96, Vol. 2, pp. 1237-1241, October 1996.

[YJ02] H. Yousefizadeh, and H. Jafarkhani, "Analytical modeling of burst loss: A study of the Gilbert model," *In Proceeding. of ACM SIGEMTRICS*, pp. 1-4, July 2002.

[YS01] P. Yuan, and A. Skoe, "Design and implementation of scalable edge-based admission control," *Elsevier Computer Networks,* Vol. 37, pp. 507-518, April 2001.

[YSu01] B. Yener, and G. Su , "Smart box architecture: a hybrid solution for IP QoS provisioning," *Elsevier Computer Networks,* Vol. 36, No.4, pp. 357-375, July 2001.

[ZY02] C. Zhu, and O. Yang, "A comparison of active queue management algorithm using the OPNET Modeler," *IEEE Communication Magazine,* Vol. pp. 158 167, June 2002.